Univerzita Karlova v Praze
Přírodovědecká fakulta

Studijní program: Biologie
Studijní obor: Zoologie

Bc. Miroslava Loudová

Zavedení metod RAD sekvenování do výzkumu genetické struktury ježků rodu *Erinaceus*

Implemenation of the RAD sequencing methods to the population genetic studies of hedgehogs from the genus *Erinaceus*

Diplomová práce

Vedoucí práce: Mgr. Barbora Černá Bolfíková, Ph.D.
Konzultant: doc. RNDr. Pavel Hulva, Ph.D.

Praha, 2015

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 14. 8. 2015

Podpis

Poděkování

Na tomto místě bych chtěla poděkovat především mé školitelce Mgr. Barboře Černé Bolfíkové, Ph.D. za pečlivé vedení tohoto výzkumu a hlavně za cenné rady, čas, trpělivost a pevnou ruku při psaní této práce. Velký dík patří mému konzultantovi doc. RNDr. Pavlu Hulvovi, Ph.D. za konzultace při směrování této práce a zajišťování jejího uskutečnění. Dále chci poděkovat Mgr. Kristýně Eliášové za to, že mě naučila velkou část laboratorních technik potřebných pro genetický výzkum a RNDr. Lukášovi Cholevovi, Ph.D. za poskytnutí primerů pro sekvenaci na platformě Illumina. Děkuji také celému týmu genomické laboratoře v European Molecular Biology Laboratory (EMBL) v Heidelbergu za pomoc při přípravě knihovny pro sekvenování pomocí metody tzv. Next-generation sequencing. V neposlední řadě chci poděkovat všem, co se podíleli na sběru materiálu. Nejvíce však vděčím za podporu své rodině a kamarádům, bez kterých bych nikdy nemohla dojít až sem.

Tento projekt byl podpořen Grantovou agenturou Univerzity Karlovy (číslo projektu 702214).

Obsah

Abstrakt

Ježci rodu *Erinaceus* představují důležitý modelový organismus pro studium postglaciální rekolonizace Evropy a procesů probíhajících v místech sekundárního kontaktu jejich areálů. V této práci bylo použito celkem pět jedinců ježka východního (*Erinaceus roumanicus*), čtyři jedinci ježka západního (*Erinaceus europaeus*) a jeden předpokládaný hybrid. Geografická distribuce jedinců využitých v analýze pokrývá oblast střední Evropy, nicméně v dalším výzkumu bude tento dataset dále rozšířen na celou oblast západního palearktu. Hlavním cílem práce bylo zavedení nové metody do výzkumu ježků, díky níž je možné zmapovat celoareálovou populačně-genomickou strukturu rodu *Erinaceus* v západním palearktu. Metodou RADSeq (Restriction-site associated DNA sequencing) lze získat polymorfní markery, např. právě námi použité SNPs (Single Nucleotide Polymorphisms), napříč genomem. V této práci bylo analyzováno celkem 16382 pozic SNPs. Za použití binárního souboru dat, označujícího přítomnost a nepřítomnost SNP u jednotlivých druhů, nebyly plně potvrzeny hypotézy vyslovené na základě analýz klasických genetických markerů z předchozích studií. V dalším výzkumu bude potřeba ověřit možný výskyt artefaktů vzniklých při bioinformaticky náročné analýze genomických dat. Na druhou stranu práce založené na klasické genetice mají značné limity, např. malý počet jaderných markerů (v předchozích studiích u ježků pokrytí méně než 20% chromozomů jedním mikrosatelitovým markerem) nebo specifika mitochondriálních lokusů (absence rekombinace, mitochondrial capture aj.). Genomické markery naznačují existenci komplikované populačně genomické architektury u zkoumaného druhu a možnost existence hybridního roje v kontaktní zóně dvou evropských ježků rodu Erinaceus.

Klíčová slova:

Erinaceus, genomika, populační genetika, RADSeq, SNP

Abstract

Hedgehogs from the genus *Erinaceus* are an important model organism for studying the postglacial recolonisation of Europe and the processes that take place in the secondary contact zones of their areas of distribution. In this study, five individuals of white-breasted hedgehog (*Erinaceus roumanicus*), four individuals of western hedgehog (*Erinaceus europaeus)* and one estimated hybrid were analysed. Geographical distribution of individuals used in the study covers the region of the Central Europe, however in the further research expansion of analsysed individuals will be needed and the whole Palearct should be sampled. The main goal was to implement novel methods in research of hedgehogs, which will enable to map the population-genomic structure of the genus Erinaceus in western Palearct. The method RADSeq (Restriction site associated DNA sequencing) enables to obtain polymorphic markers, e.g., SNPs which we used (Single Nucleotide Polymorphisms) across the genome. In this work it was analyzed 16382 SNPs. Using the binary data which indicates the presence and absence of SNPs for each species, hypotheses raised under classical analyzes of genetic markers from previous studies have not been fully confirmed. In further research it will be necessary to verify possible occurrence of biases connected with complicated matter of bioinformatic analysis in genomic data. On the other hand, studies based on classical genetics have considerable limitations, for example small amount of nuclear loci (in previous studies in hedgehogs, less than 20% of chromosomes was covered, in each case by one microsatellite marker) or characteristics of mitochondrial loci (absence of recombination, mitochondrial capture etc.). Genomic markers suggest existence of complicated population genomic architecture in investigated species and possible occurrence of hybrid swarm in the contact zone of two European hedgehogs of the genus *Erinaceus*.

Key words:

Erinaceus, genomics, population genetics, RADSeq, SNP

1 Úvod

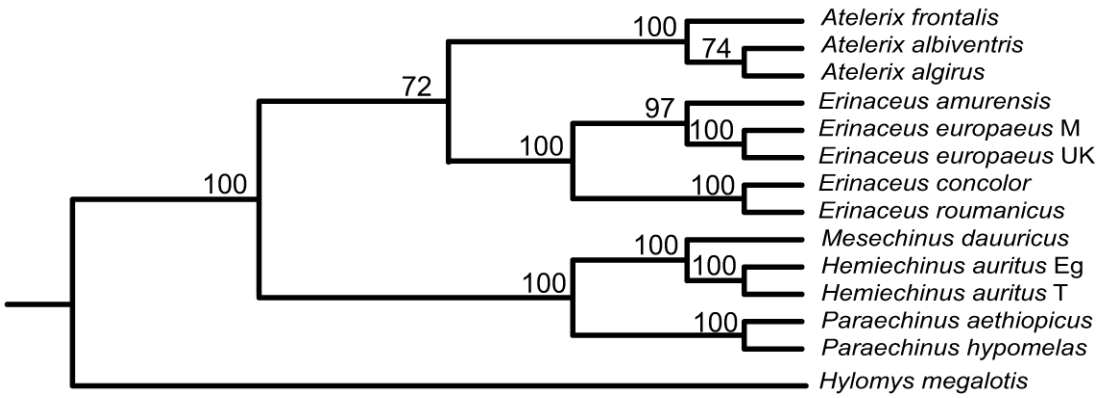
1.1 Cíle práce

Cílem této diplomové práce je zavedení nové metody do výzkumu fylogeografie a populační struktury ježků r. *Erinaceus*. Konkrétně se jedná o metodu RADSeq (Restriction-site associated DNA sequencing), pomocí které lze navrhnout polymorfní markery - SNPs (Single Nucleotide Polymorphisms).

1.2 Čeleď Erinaceidae

Čeleď Erinaceidae (ježkovití) se v současné době, spolu se Soricidae (rejskovití), Talpidae (krtkovití) a Solenodontidae (štětinatcovití), řadí do linie Laurasiatheria a řádu Eulipotyphla (Madsen *et al.* 2001; Murphy *et al.* 2001a; Murphy *et al.* 2001b; Roca *et al.* 2004). Do rodu *Erinaceus* náleží čtyři druhy: *Erinaceus europaeus* (ježek západní; Linnaeus, 1758), *E. roumanicus* (j. východní; Barrett-Hamilton, 1900), *E. concolor* (j. maloasijský; Martin, 1837) a *E. amurensis* (j. amurský; Schrenk, 1859). Všechny druhy se přirozeně vyskytují pouze v palearktické oblasti (IUCN 2015).

Až v nedávné době byl na základě molekulárních dat (např. práce Berggren *et al.* 2005; Seddon *et al.* 2002; Seddon *et al.* 2001) rozdělen druh *Erinaceus concolor* na dva různé druhy, dříve označované za poddruhy. Původní název *E. concolor* byl podle pravidel nomenklatury ponechán druhu obývajícímu Malou Asii (typová lokalita leží v asijské části Turecka). Linii žijící v Evropě, původně nazývané *Erinaceus concolor roumanicus*, připadlo jméno *Erinaceus roumanicus*. *Erinaceus roumanicus* a *Erinaceus concolor* jsou si tedy blízce příbuzní. *Erinaceus europaeus* pak podle nejnovější molekulární studie klastruje s druhem *Erinaceus amurensis*. Nejnovější fylogenetický strom rekonstruující evoluční historii podčeledi Erinaceinae je zobrazen na Obr. 1. (Bannikova *et al.* 2014).



Obr. 1. Fylogenetický strom znázorňující příbuzenské vztahy v rámci podčeledi Erinaceinae, vytvořený pomocí Bayesiánské koalescence v programu BEAST (Bannikova *et al.* 2014).

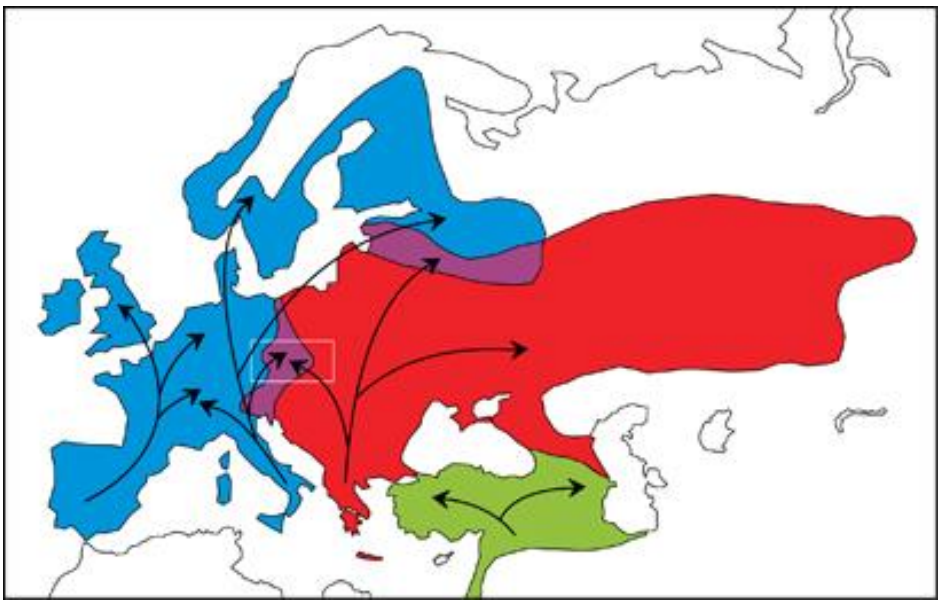
Doba divergence mezi jednotlivými druhy ovšem stále není jednoznačně vyřešena. Ve starší práci využívající analýzy cytochromu *b*, kde byl druh *E. roumanicus* ještě součástí *E. concolor*, vyšel odhad rozdělení *E. europaeus* a *E. concolor* na období před 5,8 miliony let, tedy před-pliocénní (Santucci *et al.* 1998). O tři roky později byl odhad radiace na základě cytochromu *b* a kontrolní oblasti mitochondriální DNA určen do období před 3,2 - 4,5 miliony lety do pliocénu (Seddon *et al.* 2001). Rozdělení *E. roumanicus a E. concolor* (tč. dvou linií *E. concolor*) je v prvním případě kladeno do doby před 3 miliony let, v druhém pak 1,7 - 2,2 milionů let. Studie cytochromu *b*, 12S ribozomální RNA a 5 jaderných lokusů pak stanovila rozdělení 1. linie (*E. europaeus* a *E. amurensis*) a 2. linie (*E. roumanicus* a *E. concolor*) do období 1 - 2,2 milionů let, tedy až do pleistocénu. Divergence *E. roumanicus* a *E. concolor* pak vycházela před 0,4 – 1,4 milionem let (Bannikova *et al.* 2014).

1.3 Rozšíření, ostrovy

1.3.1 Rozšíření, zóna sympatrie, hybridizace

Druh *Erinaceus europaeus* se přirozeně vyskytuje v západní Evropě. V oblasti střední Evropy na něj navazuje *E. roumanicus*, který je rozšířen až po Moskvu a západní Sibiř. *Erinaceus concolor* žije v Malé Asii. Detailněji je rozšíření těchto tří druhů znázorněno na Obr. 2. Výskyt naprosté většiny evropských druhů byl ovlivněn střídáním glaciálů a interglaciálů během kvartéru, kdy mohlo docházet k alopatrické speciaci (= vznik druhů v geograficky oddělených oblastech) v refugiích (Hewitt 2000). Dnešní rozšíření ježků v Evropě je dáno postglaciální expanzí z refugií na jihoevropských poloostrovech – z Iberského, Apeninského (pro *E. europaeus*) a Balkánského refugia (pro *E. roumanicus*) (Hewitt 1999). Zatímco *E. concolor*

nepřekonal bariéry tvořené pohořím Kavkaz a úžinami Bospor a Dardanely a do Evropy se nerozšířil (Aulagnier *et al.* 2009). Kontaktní zóna druhů *E. europaeus* a *E. roumanicus* ve střední Evropě vznikla pravděpodobně během neolitického odlesňování krajiny (Bolfikova & Hulva 2012), zatímco východoevropská kontaktní zóna až později, kdy se sem rozšířil *E. europaeus* právě ze střední Evropy (Seddon *et al.* 2001).



Obr. 2. Mapa výskytu západopalearktických ježků r. *Erinaceus*. Modrá barva odpovídá druhu *Erinaceus europaeus*, červená *E. roumanicus*, zelená *E. concolor* a fialová pak sympatrické zóně *E. europaeus* a *E. roumanicus*. Šipky naznačují pravděpodobný směr kolonizace severnějších částí pevniny z refugií po poslední době ledové (Bolfikova & Hulva 2012).

Mezi druhy je také patrná ekologická odlišnost. Areál *E. roumanicus* tvoří spíše stepní a lesostepní biomy, v České republice především nížiny. Přesto existují oblasti syntopického výskytu obou druhů (= výskytu na stejné lokalitě), například Praha. Z demografického hlediska byla populace *E. europaeus* v České republice shledána stabilní, zatímco populace *E. roumanicus* mírně rostoucí. U jedinců *E. roumanicus* byla zaznamenána vyšší míra genového toku a disperze méně závislá na pohlaví. Disperzi toho druhu tedy pravděpodobně ovlivňuje menší počet zdrojů v otevřené krajině (oproti lesům, kde se převážně vyskytuje *E. europaeus*). Dále je s velkou pravděpodobností ovlivněna menší koncentrací jedinců v krajině a následně větší potřebou hledání partnera pro páření. Při vzorkování v České

republice bylo shromážděno třikrát méně jedinců *E. roumanicus* než *E. europaeus*, což může svědčit právě o menší početnosti populace *E. roumanicus*, případně o skrytějším způsobu života (Bolfikova & Hulva 2012). V zóně sympatrie (= oblast překryvu areálů různých druhů) byl dále mezi druhy zjištěn rozdíl v prevalenci několika druhů střevních parazitů a je tedy možné, že se částečně liší i jejich potravní nároky (Pfaeffle *et al.* 2014).

Křížení těchto dvou druhů ježků v zajetí má značná omezení. Poprvé dosáhl F_1 generace Herter (1965) a později se o totéž pokusili Poduschka & Poduschka (1983, podle Reeve 1994). Těm se po několikaletém úsilí podařilo odchovat dokonce F_2 generaci, ovšem zpětné křížení hybridů F_1 generace bylo úspěšné pouze s rodičovským druhem *Erinaceus roumanicus*. Zajímavá je situace v zónách sympatrie *E. europaeus* a *E. roumanicus*. Ve střední Evropě nebyla hybridizace pomocí alozymových dat odhalena (Suchentrunk *et al.* 1998). I ve studii zahrnující analýzu kontrolní oblasti mitochondriální DNA a mikrosatelitů došli autoři k závěru, že zde recentně k hybridizaci ve volné přírodě nedochází, nebo je její frekvence velmi nízká (Bolfikova & Hulva 2012). Oproti tomu v sympatrické zóně ve východní Evropě, konkrétně v Moskevském regionu, byl jeden z pěti studovaných jedinců určen jako hybrid (Bogdanov *et al.* 2009). Vzhledem ke kratšímu trvání kontaktu obou druhů ve východní Evropě je tedy možné, že zde zatím mezi druhy *E. europaeus* a *E. roumanicus* nedošlo k úplnému vytvoření reprodukčně-izolačních mechanismů. Vzorek populace v této studii byl ovšem velmi malý, a je třeba ji doplnit podrobnějším výzkumem. Naopak ve starší středoevropské kontaktní zóně se reprodukčně-izolační mechanismy uplatňují. Lze tedy předpokládat, že hybridizace v minulosti probíhala i ve střední Evropě, ale selekce proti hybridům zde vedla právě k vytvoření reprodukčně-izolačních mechanismů.

1.3.2 Invaze a ostrovy

Biologická invaze je v dnešní době chápána jako jedna z nejčastějších příčin extinkce organismů na Zemi. V počtu tzv. invazních druhů rostliny výrazně vynikají nad živočichy, z mnoha příkladů uvedu bolševník velkolepý (*Heracleum mantegazzianum*) a akácii dlouholistou (*Acacia longifolia*) (GISD 2015; Vitousek *et al.* 1996). Jako invazní druhy savců si většinou představujeme zástupce hlodavců (Rodentia), například potkana (*Rattus norvegicus*) nebo krysu obecnou (*Rattus rattus*). Málokdo si však uvědomí, že dnešní rozšíření ježků není zdaleka původní, ale zahrnuje i oblasti, kde byl ježek záměrně vysazen. Jako už tradičně, jedná se především o ostrovy. Nejznámější případem je opakované

dovážení ježků západních (*Erinaceus europaeus*) z Velké Británie na Nový Zéland. Důvodem byl údajně čirý sentiment Evropanů, kteří se snažili připodobnit nové prostředí své domovině, a také snaha o hubení škůdců - plžů - zavlečených na ostrovy dříve (Brockie 1975; Bolfikova *et al.* 2013). Právě od jedince z Nového Zélandu je k dispozici osekvenovaný genom, který byl v této práci použit jako referenční genom pro navržení SNPs.

Až polovinu všech druhů na ostrovech mohou tvořit ty nepůvodní (Vitousek *et al.* 1996). Nezřídka tak dochází k ohrožení původní ostrovní, často endemické, bioty, například kvůli jednoduššímu složení společenstev či absenci predátorů na ostrovech (Lowe *et al.* 2000). I ježci mohou působit značné problémy původní fauně. Na ostrově South Uist u pobřeží Skotska dochází díky predaci vajec ježky západními (*Erinaceus europaeus*) k výraznému poklesu hnízdní úspěšnosti nejméně čtyř druhů ptáků z řádu Charadriiformes (dlouhokřídlí) hnízdících na zemi (Jackson 2001). Studium populační dynamiky ježků tak získává nový praktický rozměr, a to umožnění efektivního managementu druhu z důvodu záchrany původní bioty na ostrovech.

Také druh *Erinaceus roumanicus* byl v minulosti lidmi introdukován na ostrov, a to Krétu (Rackham *et al.* 1996). Existence poddruhu *E. roumanicus nesiotes* (Bate, 1906), vyskytujícím se právě na Krétě a několika dalších menších ostrovech, byla v loňském roce podpořena i analýzou mikrosatelitových dat. Krétská populace zde klastrovala odděleně od populací z Balkánského poloostrova a střední Evropy, což svědčí o jejím unikátním postavení v rámci druhu. Pouze dva jedinci z Peloponéského poloostrova byli přiřazeni k populaci z Kréty, z čehož vyplývá, že pravděpodobně pochází ze stejné zdrojové populace, případně jedna populace z druhé (Eliášová 2014).

1.4 Next-generation sequencing

Naše pracovní skupina dlouhodobě řeší otázky týkající se fylogeografie a populační genetiky ježků r. *Erinaceus*, do současné doby především za pomoci analýzy mikrosatelitů a kontrolní oblasti mitochondriální DNA. Vzhledem k limitům těchto markerů bylo při zadávání metodiky této diplomové práce přistoupeno k zavedení moderního a rychle se rozvíjejícího přístupu populační genomiky, jakým je právě RADSeq, který může poskytnout zcela nový pohled na problematiku, kterou se zabýváme. Právě kvůli významu této metody se jejímu popisu budu věnovat podrobněji.

„Next-generation sequencing" metody (= sekvenování příští generace), někdy také nazývané massively parallel sequencing (= rozsáhlé paralelní sekvenování) zažívají v poslední době nebývalý rozkvět. Oproti klasické Sangerově metodě poskytují kratší sekvence (obvykle do 400bp), jejich počet je ovšem značně vyšší, může jít až o miliony sekvencí (Reis-Filho 2009). Vyšší chybovost sekvenačních metod příští generace je kompenzována mnohonásobným pokrytím jednoho úseku DNA (tj. vysoká „coverage").

Nejdříve byly tyto moderní přístupy aplikovány na resekvenování lidského genomu (Bentley *et al.* 2008) a široké uplatnění nacházejí také v medicíně. Jako příklad mohu uvést genetickou diagnostiku rakoviny prsu (Morgan *et al.* 2010) nebo výzkum genomu viru HIV-1 (Gall *et al.* 2012). Rapidní vývoj next-gen (= next generation) sekvenování umožnil výrazné zlevnění i zrychlení sekvenování genomů. Pro srovnání, 10x prosekvenovaný lidský genom lze získat v jednom běhu přístroje a cena je až 200 000x nižší, než pomocí klasických metod, které použilo „Human Genome Sequencing Consortium" při prvním sekvenování lidského genomu, které trvalo přibližně 13 let (Reis-Filho 2009).

Next-gen sekvenování lze využít také při vývoji nových mikrosatelitových markerů, což se stává užitečným především při studiu nemodelových organismů. Tento způsob byl použit např. v práci monitorující hybridní zónu mezi strnadci ostrochvostými (*Ammodramus caudacutus*) a strnadci Nelsonovými (*Ammodramus nelsoni*) při severovýchodním pobřeží Severní Ameriky. Bylo vytipováno 12 polymorfních lokusů, s co nejmenším počtem alel, které sdílí oba druhy. Pomocí těchto markerů je možné zařadit jedince s velmi vysokou pravděpodobností do skupiny rodičovských druhů, F_1 hybridů a zpětně zkřížených jedinců.

Pouze u F_2 hybridů byla pravděpodobnost správného určení nižší, a to 76% (Kovach *et al.* 2015).

Jinou možností, jak při výzkumu hybridů použít next-gen sekvenování je navržení SNPs (Single Nucleotide Polymorphisms). SNPs jsou jednonukleotidové odchylky v sekvenci DNA mezi jedinci určitého druhu. Tento postup byl aplikován na rozlišení velbloudů dvouhrbých (*Camellus bactrianus*), velbloudů jednohrbých (*Camellus dromedarius*) a jejich kříženců. Bylo vytipováno 12 SNPs, díky kterým lze výše uvedené skupiny jednoznačně rozlišit. Tyto lze uplatnit nejen ve výzkumu recentně žijících populací, ale i u méně zachovalých archeologických vzorků. Takové vzorky lze nalézt v oblastech na Blízkém východě, kde k záměrnému křížení dlouhodobě docházelo, ale i v Evropě, kam se velbloudi dostali během války s Osmany, kteří je využívali jako dopravní prostředek (Galik *et al.* 2015; Ruiz *et al.* 2015).

1.4.1 RADSeq

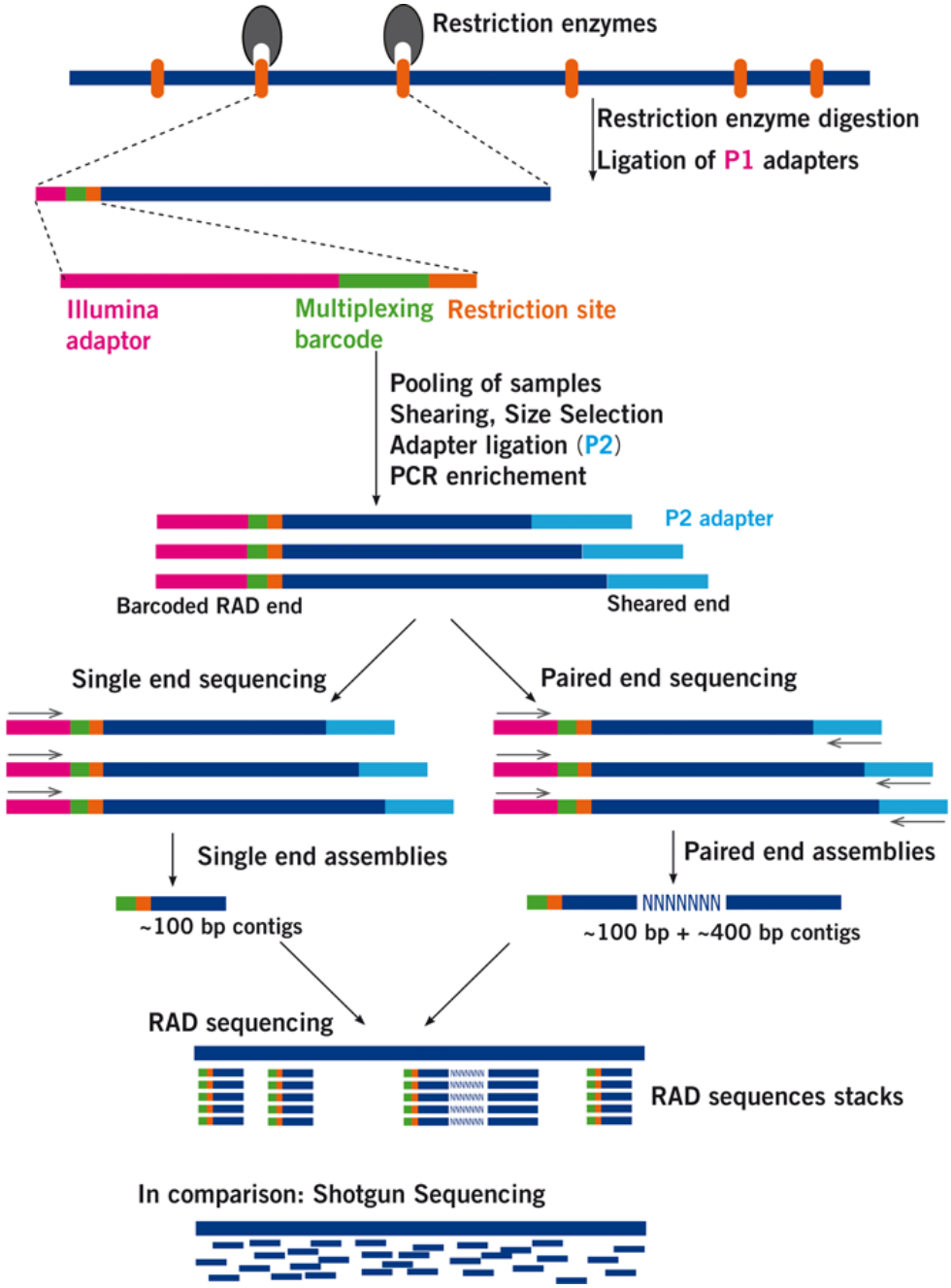
1.4.1.1 Princip

RADSeq (Restriction-site associated DNA sequencing) se řadí mezi „Next-generation sequencing“ metody. Jedná se o sekvenování oblastí genomu vybraných na základě délky po restrikčním štěpení DNA. Zprostředkovává identifikaci polymorfních markerů, např. právě SNPs. Pomocí metody RADSeq jsou objevovány SNPs v náhodných a především nekódujících oblastech genomu.

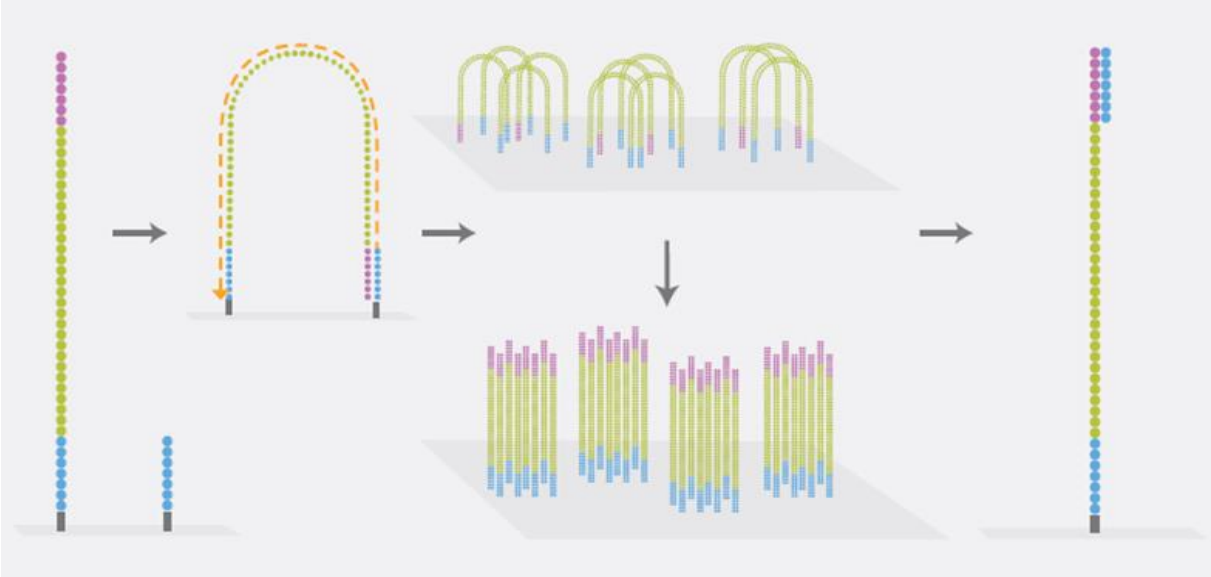
Sekvenování vyššího počtu jedinců v jednom běhu umožňuje „barcoding“ pomocí oligonukleotidů, tzv. MIDs (Multiplex Identifier Sequences) označujících sekvence DNA od jednoho jedince. Samotná sekvenace probíhá na platformě Illumina, pomocí DNA syntézy. RADSeq poskytuje až miliony sekvencí o délce 50 – 600bp (Davey & Blaxter 2010; Illumina Inc. 2014).

Průběh metody RADSeq je zobrazen na Obr. 3. Nejdříve je k DNA přidán restrikční enzym (např. SbfI), který vyhledá restrikční místa v genomu a DNA naštěpí. Různé enzymy mají v genomu různý počet restrikčních míst a výběrem určitého enzymu lze tedy ovlivnit počet vznikajících fragmentů. Existuje také metoda ddRADSeq (double digest RADSeq), ve které jsou použity dva různé restrikční enzymy současně (Peterson *et al.* 2012). V dalším kroku je

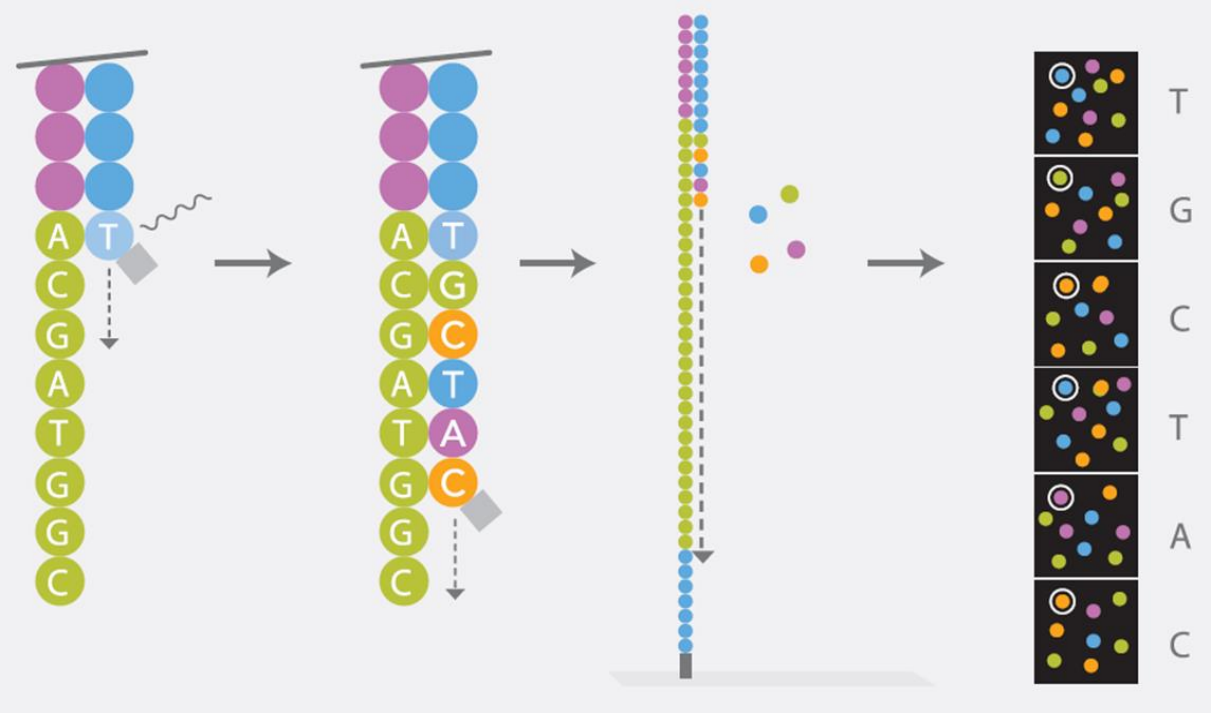
na restrikční místo ligován adaptér P1, obsahující specifický MID pro označení DNA od daného jedince. Poté je DNA od všech jedinců smíchána dohromady a ze směsi jsou vyselektovány pouze fragmenty určité velikosti. Posléze dochází k ligaci adaptéru P2 s jedním rozvětveným koncem do tvaru Y, na všechny fragmenty DNA. Tento konec se pak naváže na P2 primer pouze v případě, když je dotvořen amplifikací (= namnožením) P1 adaptéru. Tento rozvětvený konec tedy zajistí, že následná PCR reakce namnoží pouze fragmenty obsahující P1 i P2 adaptér. Jedná se o tzv. můstkovou PCR („bridge“ PCR), která je znázorněna na Obr. 4. Na speciálně upravenou skleněnou destičku (nazývanou „Flow Cell“) s předem ukotvenými primery jsou přidány fragmenty DNA. Tyto fragmenty pomocí svých adaptérů přisednou na primery umístěné na skleněné destičce. Po přidání volných nukleotidů a enzymů je podle templátu - fragmentu DNA - nasyntetizováno druhé vlákno DNA. Pomocí denaturace se pak dvouvláknový můstek rozpojí a každé vlákno zůstane přichyceno jedním koncem ke skleněné destičce. Amplifikace se opakuje, až do zaplnění skleněné destičky a dochází k vytvoření klastrů s namnoženými shodnými fragmenty DNA. Sekvenace probíhá pomocí DNA syntézy, na identické skleněné destičce (Obr. 5). Na destičku jsou vylity všechny čtyři typy nukleotidů najednou. Každý typ je označen jiným fluorescenčním barvivem a obsahuje také terminátor, který zabrání připojení většího počtu nukleotidů v jednom kroku. Po zařazení nukleotidu do syntetizovaného vlákna DNA nukleotid zasvítí a je detekován pomocí laseru. Po odstranění terminátoru se tento cyklus opakuje až do chvíle, kdy je osekvenován celý fragment. Na platformě Illumina lze využít tzv. „pair-end“ sekvenování, kdy je fragment DNA čten z jednoho i z druhého konce a tím je výrazně usnadněna „assembly“ (= skládání, rekonstrukce) sekvencí (Davey & Blaxter 2010; Mardis 2008; Illumina Inc. 2010).



Obr. 3. Schéma znázorňující průběh metody RADSeq. DNA je naštěpena restrikčním enzymem. Na jednotlivé fragmenty je připojen adaptér P1, obsahující MID (krátkou sekvenci oligonukleotidů) pro označení sekvencí od každého jedince. Poté je DNA smíchána, jsou vybrány fragmenty o určité délce a je připojen adaptér P2. Fragmenty obsahující P1 i P2 adaptér jsou namnoženy pomocí můstkové PCR. Následuje samotná sekvenace. Převzato z http://www.floragenex.com/rad-seq/.



Obr. 4. Můstková PCR probíhající na platformě Illumina. Fragmenty DNA jsou pomocí adaptéru připojeny k primerům ukotveným na skleněné destičce. Podle templátu je nasyntetizováno druhé vlákno DNA. Poté je můstek rozpojen a každé vlákno zůstává k destičce přichyceno jedním koncem. Celý proces se opakuje až do zaplnění destičky. Převzato z http://www.cegat.de/en/services/next-generation-sequencing/ngs-technology/#top



Obr. 5. Zobrazení sekvenování pomocí DNA syntézy na platformě Illumina. Na destičku jsou nality nukleotidy s fluorescenčním značením. Při zakomponování nukleotidu do řetězce DNA nukleotid zasvítí specifickou barvou, což zaznamená laser. Tímto způsobem je odečtena celá sekvence fragmentu DNA. Převzato z http://www.cegat.de/en/services/next-generation-sequencing/ngs-technology/#top

1.4.1.2 Využití RADSeq metod v populační genetice

Metoda RADSeq má v moderním biologickém výzkumu široké spektrum využití, např. ve fylogenomice, fylogeografii, a populační genomice. RADSeq je samozřejmě využíváno i při studiu rostlin, a to i ekonomicky významných (např. Wu *et al.* 2014), vzhledem k zaměření této práce se jimi ale zde nebudu blíže zabývat.

Jednou z prvních zásadních studií, využívajících přístupu RADSeq na nemodelových druzích živočichů, je bezesporu práce o koljuškách tříostných (*Gasterosteus aculeatus*). Výzkum byl proveden na třech sladkovodních a dvou oceánských populacích v oblasti Aljašky, kdy z každé populace bylo použito 20 jedinců. Další výsledky byly získány pomocí křížení ryb z těchto pěti populací v laboratorních podmínkách, a to zařazením F_2 generace do výzkumu. Byly sledovány znaky objevující se u sladkovodních populací: redukce laterálních štítků a pánevních struktur. Byly vytipovány kandidátní geny, které s jejich redukcí souvisí. Mimo

lokusu EDA (pro gen ectodysplasinu), který byl detekován pomocí QTL (Quantitative trait loci) v dřívější studii (Colosimo *et al.* 2005), byly nalezeny i další regiony v genomu, které souvisí s fenotypovou evolucí těchto dvou znaků. Pomocí 45 789 SNPs a populačně-genetických statistik bylo zjištěno, že mezi oceánskými populacemi probíhá výrazný genový tok a nejsou od sebe vzájemně izolovány. Lze tedy hovořit o jedné velké populaci, která dala nezávisle a opakovaně vzniknout populacím sladkovodním (Hohenlohe *et al.* 2010).

Mechanismy introgrese a genetické diferenciace byly zkoumány u pěvců pipulek bělolímcových (*Manacus candei*) a pipulek zlatokrkých (*Manacus vitellinus*) ve Střední Americe. Do studie byla zahrnuta především hybridní zóna ležící v Panamě a analyzováno bylo 59 100 SNPs. Bylo zjištěno, že lokusy spojené s vyšší mírou genetické diferenciace a introgrese jsou roztroušeny po celém genomu a nejsou tedy soustředěny jen v několika málo oblastech. Mezi introgresí a genetickou diferenciací byla nalezena částečná pozitivní korelace a zdá se tedy, že reprodukčně-izolační mechanismy a lokální adaptace jsou u pipulek z genetického hlediska propojené (Parchman *et al.* 2013).

Využití ve fylogeografii dokumentuje práce o komárech druhu *Wyeomyia smithii* v západní části Severní Ameriky. Přístup RADSeq potvrdil rozdělení datasetu na severní a jižní část, které bylo navrženo dříve pomocí analýzy části mitochondriální DNA, a to genu pro COI (cytochrom oxidáza I). Dále RADSeq umožnil jemnější rozdělení těchto dvou skupin, každou na dvě podskupiny. Nejsevernější podskupina zahrnuje i populace, které vznikly po ústupu pevninského ledovce, přibližně před 20 000 lety. Původ těchto populací byl určen do jižní části Apalačského pohoří (Emerson *et al.* 2010).

Populační genomika může být velmi nápomocná při hledání odpovědí o skrytě žijících organismech, kdy není možné učinit závěry například o migraci pouze přímým pozorováním. Takovým případem je studie o koníčkovi vzpřímeném (*Hippocampus erectus*) provedená při západním pobřeží Atlantiku. V temperátní provincii Virginia byli koníčci v zimním období pozorováni jen velmi zřídka. Nabízela se tedy možnost, že sem v teplých měsících migrují jedinci z jižněji položených provincií, kteří po sezóně hynou. Pomocí 11 708 SNPs byly identifikovány tři oddělené subpopulace a jednu z nich tvořili jedinci právě z provincie Virginia. Virginská populace byla jasně vymezená od ostatních, což svědčí o dlouhodobé izolaci a omezení genového toku s jižními populacemi. Nízká četnost pozorování v zimních

měsících tak bude dána spíše skrytějším způsobem života a pravděpodobně zimováním na volném moři, nikoli úhynem (Boehm *et al.* 2015).

Velmi zajímavé výsledky často poskytují studie zabývající se ostrovy. Na ostrovech São Tomé a Príncipe v Guinejském zálivu se vyskytují endemické druhy žab, rákosnička ostrovní (*Hyperolius thomensis*) a rákosnička Mollerova (*Hyperolius molleri*). Na mladším, ale větším ostrově São Tomé se vyskytují oba druhy, na ostrově Príncipe pouze rákosnička Mollerova. Pomocí analýzy mitochondriální DNA byly odhaleny 3 haplotypy: jeden pro rákosničky Mollerovy žijící na São Tomé, jeden pro žijící na Príncipe a jeden pro rákosničky ostrovní ze São Tomé. Dále byl použit přístup ddRADSeq pro navržení SNPs, po jejichž analýze došli autoři k závěru, že druhy mají monofyletický původ a vznikly alopatricky na ostrově São Tomé. Odtud se rákosnička Mollerova dostala na ostrov Príncipe , kde vznikla nová populace, která je pravděpodobně díky efektu zakladatele či genetickému driftu velmi odlišná od zdrojové populace. Ke genovému toku mezi ostrovy nedochází. Na ostrově São Tomé došlo nejspíše díky zemědělské činnosti člověka k rozšíření areálu rákosničky Mollerovy a tím ke vzniku sekundárního kontaktu obou druhů, což zapříčinilo vznik hybridní zóny (Bell *et al.* 2015).

RADSeq najde uplatnění také při studiu populační dynamiky. U čmeláků druhu *Bombus impatiens* ve východní části USA byla za pomoci analýzy mikrosatelitů nalezena výrazně vyšší genetická diverzita než u druhu *Bombus pensylvanicus* (Lozier *et al.* 2011). Po použití SNPs, pro které vyšla genetická diverzita obou druhů výrazně podobnější, se dříve publikovaný závěr, že je *Bombus pensylvanicus* ustupujícím druhem, nejeví být správným. Naopak se zdá, že oba druhy prošly v relativně nedávném čase populační expanzí. A tak ztráta genetické diverzity zjištěná pomocí mikrosatelitů nemusí být dána snížením velikosti populace, ale naopak jejím náhlým nedávným nárůstem (Lozier 2014).

1.4.1.3 Shrnutí

Jednou z hlavních výhod metody RADSeq je nesporně možnost vývoje velkého množství markerů, pokrývajících reprezentativně celý genom, zároveň je však objem těchto dat menší než u celogenomové sekvence, což usnadňuje jejich zpracování. Nejčastěji je jedná právě o SNPs, které byly využity i v této diplomové práci. Současně je možné tuto metodu aplikovat na nemodelové organismy (Emerson *et al.* 2010). Nevýhodou této metody je výrazně

náročnější bioinformatické zpracování dat. Ačkoli již vyšla řada prací využívajících přístup RADSeq v populační genetice, neexistuje zatím žádná studie využívající tuto metodu pro výzkum savců.

2 Materiál a metody:

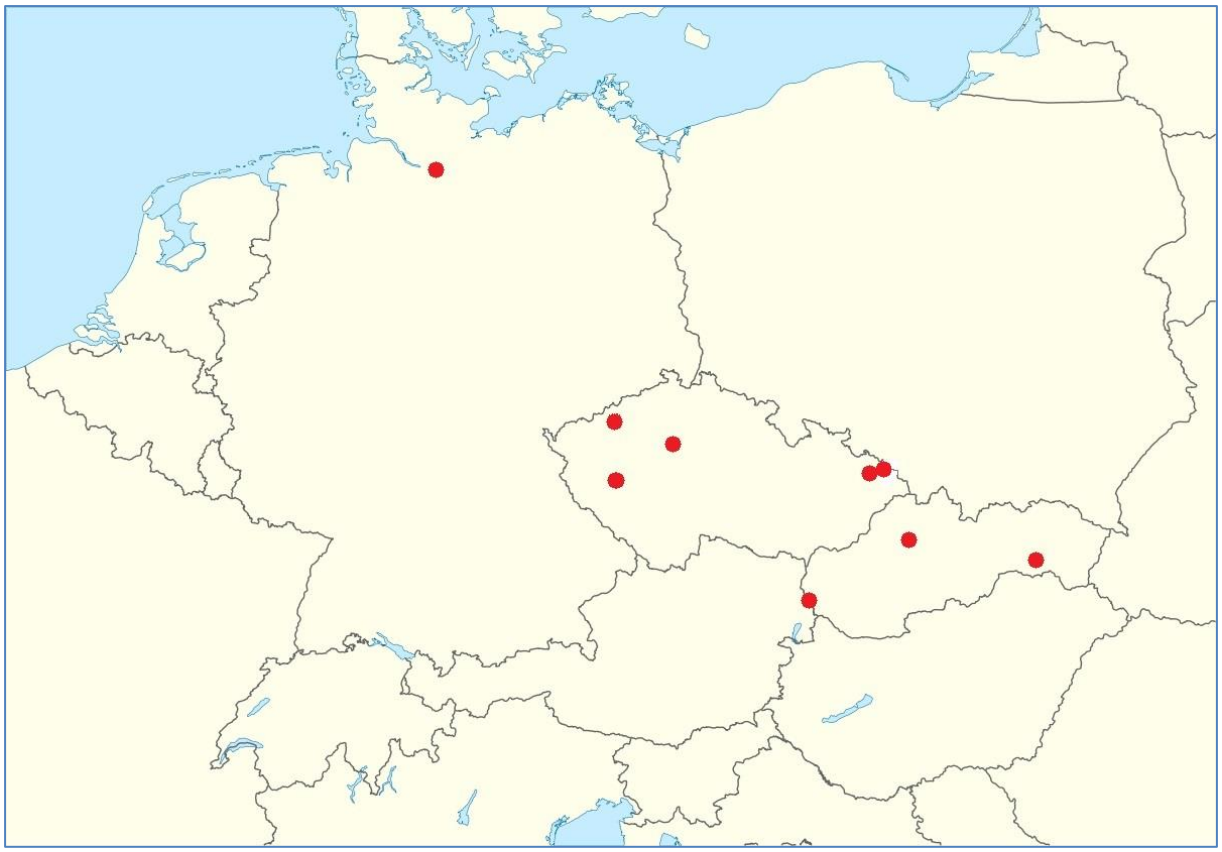
2.1 Vzorky

Vzorky pochází z oblasti západního palearktu, konkrétně ze střední Evropy, a převážná část materiálu pro tuto molekulární studii byla získána z jedinců uhynulých na silnicích. Většina vzorků již byla použita v předchozích fylogeografických studiích naší výzkumné skupiny, a to za použití analýzy mikrosatelitů.

Celkem bylo do analýzy zahrnuto 10 jedinců, jejichž distribuce reprezentativně pokrývá oblast výskytu r. *Erinaceus* ve střední Evropě, zahrnující i sympatrickou zónu druhů *Erinaceus europaeus* a *Erinaceus roumanicus*. Konkrétně byli zařazeni jedinci z těchto států: Německo, Česká republika a Slovensko. Byl použit také vzorek sk27, který byl dříve za pomoci analýzy 9 mikrosatelitů a části mitochondriální DNA (D loop) určen jako hybrid. D loop klastroval společně s druhem *E. europaeus*. V programu NewHybrids byl vzorek zařazen do kategorie F_2 (0,439) a do kategorie zpětného křížence F_1 s druhem *E. roumanicus* (0,553) (Černá Bolfíková 2013). Přiřazení do kategorie zpětného křížence F_1 s *E. roumanicus* odpovídá i poznatkům o křížení v zajetí, protože zpětného křížení s druhem *E. europaeus* nebylo dosaženo (Poduschka & Poduschka 1983, podle Reeve 1994).

Přesné lokality sběru jsou uvedeny v tabulce Tab. 1 a geograficky znázorněny na Obr. 6.

Genetický materiál byl uchováván v mikrozkumavkách v 96% ethanolu při teplotě -20°C. Pro izolaci DNA byly použity tkáně - nejčastěji sval, kůže nebo ucho.



Obr. 6. Grafické znázornění distribuce lokalit sběru jedinců použitých ve studii. Vytvořeno pomocí
https://commons.wikimedia.org/wiki/File:Central_Europe_location_map.svg.

Tab. 1. Místa a data sběru jednotlivých vzorků použitých v analýze.

označení vzorku	druh	země původu	lokalita	datum sběru	GPS souřadnice		sběr
SK32	*Erinaceus roumanicus*	Slovakia	Devínská Nová Ves	2.8.2009	48,24	16,95	Noga Michal
112	*Erinaceus roumanicus*	Czech Republic	Praha Bohnice	31.5.2008	50,12	14,42	Pithartová Tereza
122	*Erinaceus europaeus*	Czech Republic	Praha Troja	24.5.2008	50,12	14,42	Pithartová Tereza
453	*Erinaceus europaeus*	Germany	Hamburg	NA	53,55	9,98	Petney Trevor, Pffafle Miriam, Skuballa Jasmin
183	*Erinaceus roumanicus*	Czech Republic	Chomutov	2008	50,45	13,4	Zoo Chomutov
211	*Erinaceus europaeus*	Czech Republic	Klimkovice	1.11.2008	50,18	18,12	ZŠ Nový Jičín
197	*Erinaceus europaeus*	Czech Republic	Spálené Poříčí	26.11.2008	49,6	13,6	ČSOP Spálené Poříčí
209	*Erinaceus roumanicus*	Czech Republic	Rychvald	9.4.2009	49,85	18,37	Bolfíková Barbora, Hulva Pavel, Janko Karel
SK24	*Erinaceus roumanicus*	Slovakia	Košice	2.7.2009	49,09	21,31	Celuch Martin
SK27	*hybrid*	Slovakia	Diviaky	10.9.2009	48,88	18,86	Celuch Martin

2.2 Izolace DNA

Izolace DNA byla provedena komerčně dodávaným kitem „DNA Blood and Tissue Kit“ od firmy Quiagen. Přiložený protokol „Purification of Total DNA from Animal Tissue“ byl upraven ve třech bodech. V druhém kroku protokolu bylo pro lyzi tkáně použito 15 µl proteinázy K, dalších 5 µl bylo přidáno po rozložení tkáně a dále inkubováno v termostatu po dobu 20 minut. V sedmém bodě protokolu bylo pro vymytí DNA použito 100 µl elučního pufru AE. Poslední krok protokolu, opakování eluce, byl vynechán.

Vyizolovaná DNA byla následně uskladněna při teplotě -20 °C.

2.3 Kontrola koncentrace, čistoty a integrity genomické DNA

Koncentrace získané DNA byla změřena na spektrofotometru ND-1000 (Nanodrop®) v Laboratoři sekvenace DNA na Přírodovědecké fakultě UK v Praze a na fluorimetru Qubit® 2.0 od firmy LifeTechnologies v European Molecular Biology Laboratory v Heidelbergu. Pomocí poměru A260/A280 byla odhadnuta čistota izolátu. Integrita DNA byla zjišťována na přístroji Bioanalyzer od firmy Agilent za použití protokolu „Agilent High Sensitivity DNA Kit“ (G2938- 90321). Na základě získaných výsledků byly z poolu většího množství vzorků vybrány ty s dostatečnou kvalitou DNA pro následné genomické analýzy.

2.4 D loop

2.4.1 PCR

Pro amplifikaci části sekvence mitochondiální DNA, D loopu, byly použity primery ProL-He (5‘-ATACTCCTACCATCAACACCCAAAG-3‘) a DLH-He (5‘-TCCTGAAGAAAGAACCAGATGTC-3‘). PCR reakce byla provedena v termocykleru iCyclerTM Thermal Cycler (Bio–RAD). Složení reakční směsi je uvedeno v Tab. 2 a program reakce v Tab. 3.

Tab. 2. Složení a koncentrace reakční směsi pro PCR reakci D loopu.

	c	V [µl]
PCR master mix		12,5
Primer DLH-He	10 µM	1
Primer ProL-He	10 µM	1
H_2O		8,5
DNA		2
celkem		25

Tab. 3. Program termocykleru pro PCR reakci D loopu.

cyklus	T [°C]	t [min]
1 (1x)	94	3
2 (29x)	94	1
2	56	1
2	72	1
3 (1x)	72	4
4 (1x)	12	∞

2.4.2 Přečištění

Přečištění PCR produktu bylo provedeno pomocí komerčně dodávaného kitu „Qiaquick PCR purification kit“ od firmy Quiagen. Protokol od výrobce „PCR purification spin protocol“ byl modifikován pouze v posledním kroku, kdy bylo pro vymytí namnoženého úseku mitochondriální DNA z membrány použito 30 µl pufru AE.

2.4.3 Sekvenace mitochondriálního markeru Sangerovou metodou

Sekvenační analýza proběhla na sekvenátoru 3130 Genetic Analyzer v Laboratoři sekvenace DNA Přírodovědecké fakulty UK.

Tab. 1. Složení reakční směsi pro sekvenaci D loopu.

	c	V [µl]
primer DLH-He	10 µM	0,5
H_2O		5,5
PCR produkt		2
celkem		8

Jedinci byli díky získaným sekvencím za použití databáze GenBank přiřazeni k jednotlivým druhům.

2.5 RADSeq

2.5.1 Odhad počtu restrikčních míst v genomu *E. europaeus* a teoretická optimalizace metody

Nejdříve byl pomocí vhodné aplikace (RADtag counter from GenePool, Edinburgh; viz Appendix 1) odhadnut počet restrikčních míst v genomu *E. europaeus* pro jednotlivé restrikční enzymy a pomocí získaných informací byly optimalizovány parametry metody.

K tomu byly použity informace o genomu *E. europaeus* z GenBank (proporce 0.42 GC, velikost 2708 MB).

Při plánovaném výstupu 100 miliónů čtení, sekvenování 25 vzorků při „coverage“ 30 se jako nejvhodnější enzym ukázala restriktáza SbfI. Při získaném odhadu 19 533 restrikčních míst v genomu zkoumaného druhu je při zvolených parametrech možné sekvenovat v jednom běhu přístroje.

2.5.2 Příprava knihovny

Vyizolovaná DNA byla sekvenována za použití přístupu RADseq, na platformě Illumina (typ HiSeq2000, paired end, 100bp reads) v European Molecular Biology Laboratory v Heidelbergu (Německo). K štěpení DNA byl použit restrikční enzym SbfI, sekvence restrikčního místa je CCTGCA*GG (* znázorňuje místo štěpení). Sekvence P1 adaptérů jsou uvedeny v Tab. 5 a sekvence adaptéru P2 je následující:

P2-FORWARD

5´- /5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA-3´

P2-REVERSE

5´CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T -3´.

Tab. 5. Sekvence P1 adaptérů ligovaných na jednotlivé vzorky. Sekvence pěti nukleotidů v názvu P1 adaptéru odpovídá sekvenci MID.

označení vzorku	**název P1 adaptéru**	**sekvence P1 adaptéru**
SK32	P1-FOR-CGGCG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGCGTGC*A
112	P1-FOR-CGTAT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTATTGC*A
122	P1-FOR-CTAGG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAGGTGC*A
453	P1-FOR-CTTCC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCCTGC*A
183	P1-FOR-GAAGC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAGCTGC*A
211	P1-FOR-GACTA	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGACTATGC*A
197	P1-FOR-GAGAT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGATTGC*A
209	P1-FOR-GATCG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCGTGC*A
SK24	P1-FOR-GCATT	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATTTGC*A
SK27	P1-FOR-GCCGG	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCGGTGC*A
SK32	P1-REV-CGGCG	/5Phos/CGCCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
112	P1-REV-CGTAT	/5Phos/ATACGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
122	P1-REV-CTAGG	/5Phos/CCTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
453	P1-REV-CTGAA	/5Phos/TTCAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
183	P1-REV-CTTCC	/5Phos/GGAAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
211	P1-REV-GAAGC	/5Phos/GCTTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
197	P1-REV-GACTA	/5Phos/TAGTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
209	P1-REV-GAGAT	/5Phos/ATCTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
SK24	P1-REV-GATCG	/5Phos/CGATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T
SK27	P1-REV-GCATT	/5Phos/AATGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T

Pro přípravu knihovny byl použit protokol „Sequenced RAD Markers for Rapid SNP Discovery and Genetic Mapping“ (Etter 2008, modifikace Choleva; viz Appendix 2) s těmito změnami:

V bodě 3.2.1 bylo použito 500ng DNA v objemu 44,5 µl.
V bodě 3.2.2 probíhala inaktivace restrikčního enzymu při 80°C po dobu 20 min.
V bodě 3.3.2 byla směs inkubována při pokojové teplotě po dobu 40 min.
Kroky 3.5, 3.6, 3.7, 3.8, 3.9 a 3.10 byly nahrazeny jinými protokoly:

Pro přečištění DNA byl použit protokol „Agencourt® AMPure® XP PCR Purification (000387v001, 96 Well Format)“ s následujícími změnami:

V bodě 4 byla směs ponechána v magnetickém stojanu 5 min.
V bodě 6 byl použit 80% ethanol.
V bodě 7 bylo pro vymytí použito 50 µl H_2O.
V bodě 8 byl roztok ponechán v magnetickém stojanu 5 min.

Sonikace byla provedena na ultrasonifikátoru S2 od firmy Covaris®.

Kvalita DNA a velikost fragmentů byla zjišťována na přístroji Bioanalyzer od firmy Agilent za použití protokolu „Agilent High Sensitivity DNA Kit Guide“ (G2938- 90321).

Pro opravu konců DNA, ligaci adaptoru P2, přečištění a výběr fragmentů podle velikosti, namnožení DNA pomocí PCR a následné přečištění byl použit protokol „NEBNext® Ultra™ DNA Library Prep Kit for Illumina®“ (NEB #E7370S/L). Protokol byl modifikován v těchto bodech:

V bodě 1.2.3 byla směs ponechána v termocykleru po dobu 20 min.
Bod 1.2.4 a 1.2.5 byl vynechán.
V bodě 1.3A.2 bylo použito 16,5 µl H_2O.
V bodě 1.3A.3 bylo přidáno 35 µl AMPure XP beads (korálků).
V bodě 1.3A.6 bylo použito 15 µl AMPure XP beads.
V bodě 1.3A.11 bylo použito 20 µl H_2O.
V bodě 1.3A.12 bylo přeneseno 20 µl roztoku.
Bod 1.3B byl vynechán.

V bodě 1.4A.1 byly použity 4 µl DNA fragmentů, 25 µl NEBNext HighFidelity 2x PCR Master Mixu, 1 µl 25 µM Index Primeru, 1 µl 25 µM Universal PCR Primeru a 19 µl H_2O. Program pro PCR reakci je uveden v Tab. 6.

Tab. 6. Program termocykleru pro PCR reakci.

cyklus	T [°C]	t [s]
1 (1x)	98	30
2 (16x)	98	10
2	65	30
2	72	30
3 (1x)	72	300
4 (1x)	10	∞

Bod 1.4B a 1.4C byl vynechán.
V bodě 1.5.8 bylo použito 20 µl H_2O a směs byla 3 min inkubována při pokojové teplotě.

Po namnožení fragmentů byla koncentrace DNA opět změřena na fluorimetru Qubit® 2.0. Kvalita DNA a délka fragmentů byla zkontrolována na přístroji Bioanalyzer za použití protokolu „Agilent DNA 1000 Kit Guide“ (G2938-90014).

Následovala samotná sekvenace na přístroji HiSeq 2000 od firmy Illumina za použití protokolu „HiSeq® 2000 System User Guide“ (15011190). Bylo použito tzv. pair-end sekvenování.

2.6 Analýza dat

Bioinformatická analýza a vyhodnocení SNP probíhala ve spolupráci s informatikem, který poskytl vhodné hardwarové vybavení, které je nezbytné pro zpracování velkoobjemových dat typu next-gen sekvencí.

Data byla zpracovávána v operačním systému Bio-Linux (Field *et al.* 2006), který obsahuje předinstalované nástroje využitelné pro analýzu dat z „Next-generation“ sekvenování.

Sekvence je zaznamenána ve fastaq formátu, který obsahuje, kromě vlastní sekvence nukleotidů a popisu sekvencí, také znaky určující kvalitu („Phred Quality scores“) této sekvence. Vzorec pro výpočet „Phred quality scores“ je následující:

$$Q = -10 \log_{10} P.$$

„P“ určuje pravděpodobnost, s jakou je nukleotid určen nesprávně. Například pro Q = 20, je pravděpodobnost správného určení báze 99% (Illumina Inc. 2011; Ewing & Green 1998).

Pro zobrazení kvality „readů“ byl použit program FastQC 0.11.2 od společnosti Babraham Bioinformatics.

Pomocí softwaru FastqMcf (Aronesty 2011) byly odstraněny sekvence P1 a P2 adaptérů a „readů“ s příliš nízkou kvalitou určení bází. Jako limitní byla určena hodnota Q = 20, kdy je pravděpodobnost správného určení báze 99%. Všechny „ready“ obsahující báze o hodnotě Q < 20 byly vyřazeny.

Pro zkompletování (assembly) dat byla využita sekvence referenčního genomu *Erinaceus europaeus* s názvem „EriEur2“ osekvenovaná v květnu 2012 z jedince z Nového Zélandu (GCA000296755.1), dostupná na http://pre.ensembl.org/Erinaceus_europaeus/Info/Index. Právě z tohoto důvodu se zmiňuji o zajímavých faktech z ostrovní biogeografie ježků r. *Erinaceus* v kapitole 1.2.2 Invaze a ostrovy. Pro mapování sekvencí na referenční genom *E. europaeus* byl použit program Burrows – Wheeler Aligner 0.7.5a, algoritmus BWA-MEM (Li & Durbin 2010).

Pomocí Picard tools (Broad Institute), konkrétně nástroje SortSAM byly sekvence převedeny do binárního souboru ve formátu BAM. Poté byly za použití SAMtools 0.1.19 - rmdup (Li *et al.* 2009) rozpoznány a odstraněny PCR duplikáty (artefakty). Identifikace SNPs a „indels“ (polymorfismus inzerce - delece) proběhla pomocí SAMtools a byl vytvořen soubor ve formátu VCF. Následně byla provedena kontrola kvality „variant calls“. Anotace k referenčním SNPs genomu *E. europaeus* proběhla pomocí nástroje Variant Effect Predictor, dostupného online na www.ensembl.org.

Pro zobrazení mutací vyhodnocených jako SNPs v rámci jednotlivých „readů“ byl použit program Intergrative Genomics Viewer 2.3 (Robinson *et al.* 2011; Thorvaldsdottir *et al.* 2013).

Počty SNPs navržené pro každý vzorek a naměřené hodnoty koncentrace DNA pro Nanodrop a Qubit byly vyhodnoceny v programu STATISTICA (Woiss 2007). Pro tyto hodnoty byl vypočten Pearsonův korelační koeficient pro zjištění, zda koncentrace DNA souvisí s počtem získaných SNP.

Dataset obsahující shodné SNPs pozice pro všech 10 zkoumaných jedinců byl vytvořen formou zobrazující pouze shody (1) a neshody (0) oproti referenčnímu genomu. K tomuto kroku bylo přistoupeno z důvodu nižší výpočetní náročnosti. Protože se jedná o jednodušší soubor dat, byl zde předpoklad, že nám jeho analýza poskytne rychlejší náhled na populační strukturu.

Pro zpracování populačně genetickými a fylogenetickými analýzami bylo použito 16382 pozic.

Bayesianská analýza probíhala v programu MrBayes (Ronquist *et al.* 2003) ve dvou MCMC (Marcov Chain Monte Carlo) bězích se čtyřmi řetězci (jeden studený a tři teplé), analýza měla 80 milionů generací a zaznamenána byla každá stá generace. Prvních 25% záznamů bylo odstraněno jako burn-in (neboli zahřívací) perioda. Přestože průměrná standartní odchylka byla menší než 0.01, konvergence obou běhů byla ověřena v programu AWTY (Wilgenbusch et al. 2004).

Pro znázornění populační struktury byl použit program STRUCTURE 2.3.4 (Pritchard *et al.* 2000), který pracuje na principu Bayesiánské klastrové analýzy. Tento program umožnuje navržení rozdělení datasetu do různého počtu klastrů za použití genetických dat, počet klastrů je označen písmenem K. Následně určuje, kterému K odpovídá nejvěrohodnější rozdělení podle genetických vzdáleností jednotlivých vzorků v datasetu. Počet K byl zadán od K=2 do K=4 s burn-in 50 000 opakování MCMC a 450 000 MCMC analýzy.

3 Výsledky

Koncentrace izolované DNA byla měřena ve dvou krocích: na spektrofotometru ND-1000 (Nanodrop®) a na fluorimetru Qubit® 2.0 (LifeTechnologies). Naměřené hodnoty jsou uvedeny v Tab. 7. Průměrná hodnota kvality sekvencí byla shledána velmi dobrou, Q = 38.

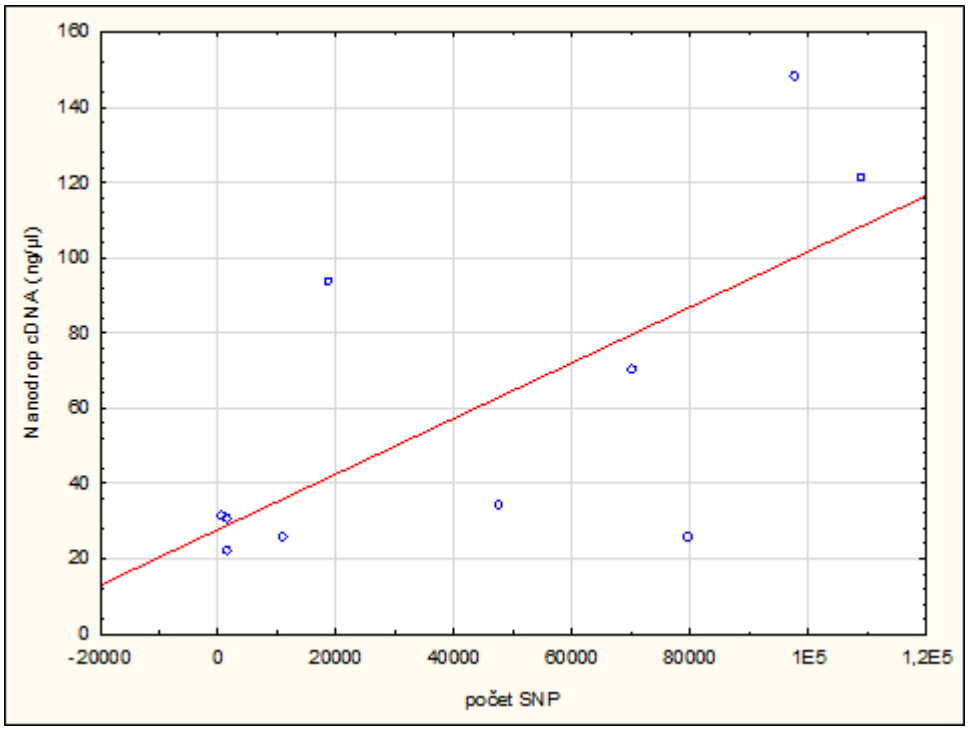
Tab. 7. Koncentrace izolované DNA z jednotlivých vzorků (NA značí neznámou hodnotu).

vzorek	**druh**	**datum sběru**	**země**	**Nanodrop cDNA (ng/µl)**	**Qubit cDNA (ng/µl)**
SK32	*Erinaceus roumanicus*	2.8.2009	Slovakia	147,8	98,4
112	*Erinaceus roumanicus*	31.5.2008	Czech Republic	121,2	90,2
122	*Erinaceus europaeus*	24.5.2008	Czech Republic	34,1	29,2
453	*Erinaceus europaeus*	NA	Germany	69,9	70,2
183	*Erinaceus roumanicus*	2008	Czech Republic	31,5	36,6
211	*Erinaceus europaeus*	1.11.2008	Czech Republic	22,3	19,5
197	*Erinaceus europaeus*	26.11.2008	Czech Republic	25,2	23,8
209	*Erinaceus roumanicus*	9.4.2009	Czech Republic	93,5	102
SK24	*Erinaceus roumanicus*	2.7.2009	Slovakia	30,3	18,7
SK27	hybrid	10.9.2009	Slovakia	25,8	26,6

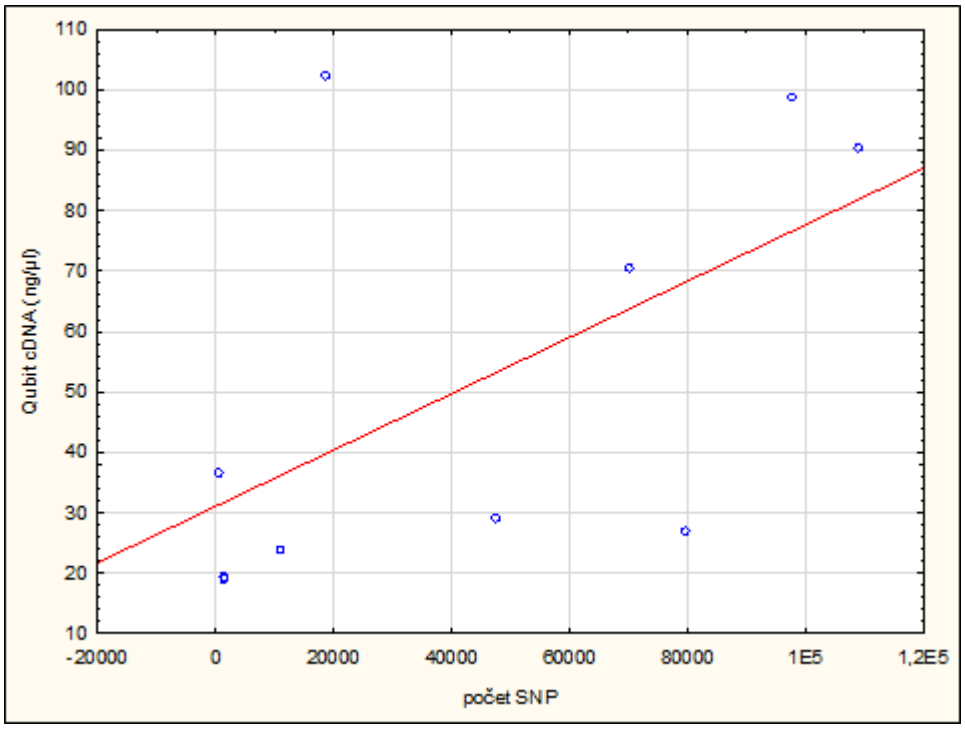
Počty SNPs navržené pro každý vzorek a naměřené hodnoty koncentrace DNA pro Nanodrop a Qubit uvádí Tab. 8. Počet SNPs signifikantně koreluje (r = 0,6867; p = 0,028) s koncentrací DNA změřené na Nanodropu, zatímco s koncentrací změřenou na Qubitu nebyla korelace potvrzena (r = 0,5715; p = 0,084). Korelace je znázorněna na Obr. 7 a 8.

Tab. 8. Počet SNPs navržených na vzorek a koncentrace izolované DNA.

		počet SNP	Qubit cDNA (ng/µl)	Nanodrop cDNA (ng/µl)
sk32	CGGCG	97710	98,4	147,8
112	CGTAT	109111	90,2	121,2
122	CTAGG	47691	29,2	34,1
453	CTTCC	70174	70,2	69,9
183	GAAGC	925	36,6	31,5
211	GACTA	1674	19,5	22,4
197	GAGAT	11199	23,8	25,2
209	GATCG	18683	102	93,5
sk24	GCATT	1641	18,7	30,3
sk27	GCCGG	79691	26,6	25,8



Obr. 9. Graf znázorňující korelaci počtu SNPs ve vzorku s naměřenou hodnotou DNA na Nanodropu.



Obr. 10. Graf znázorňující korelaci počtu SNPs ve vzorku s naměřenou hodnotou DNA na Qubitu.

Pomocí porovnání sekvence D loopu jednotlivých vzorků s databází Genebank bylo ověřeno přiřazení vzorků k danému druhu.

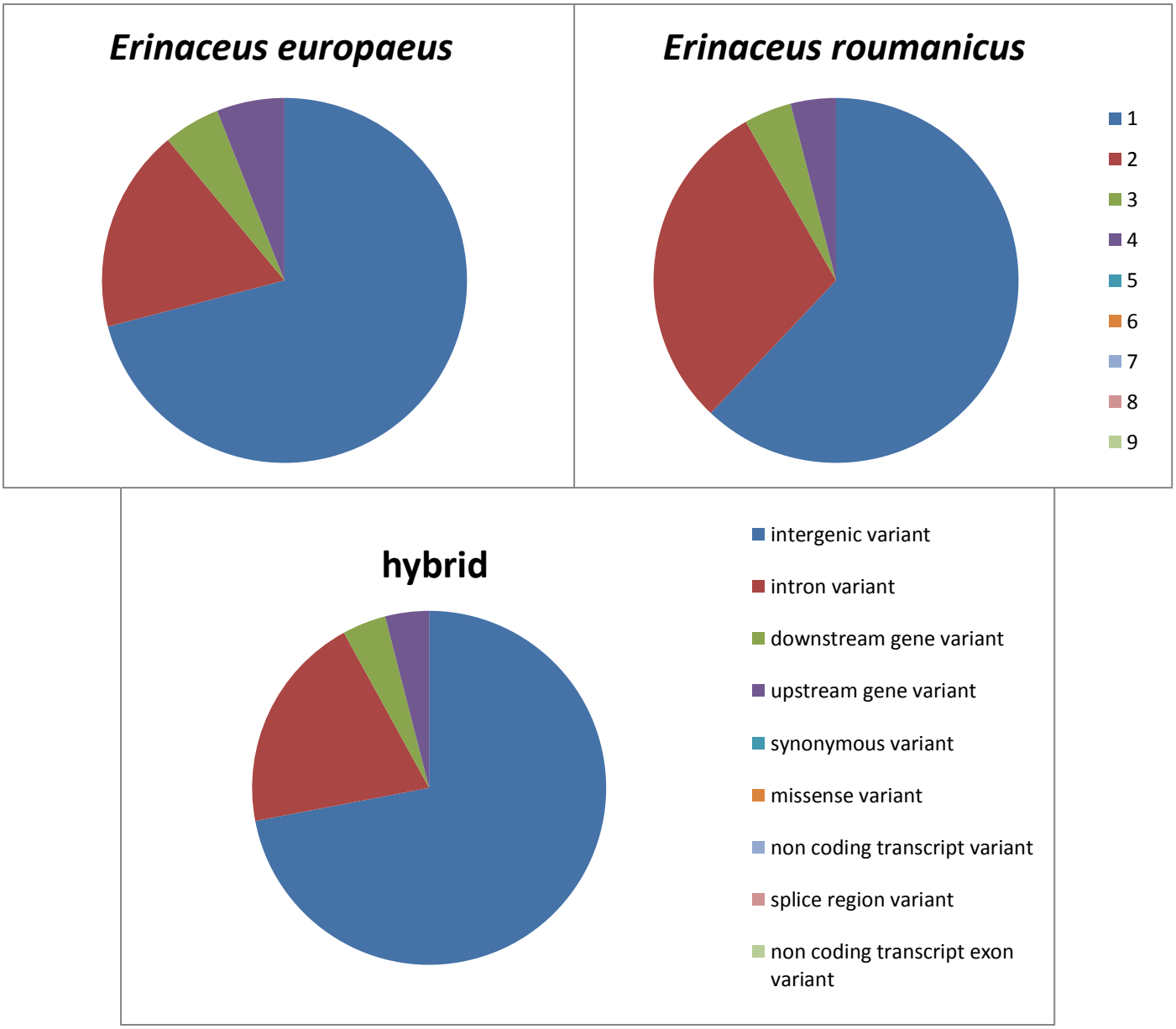
Srovnání výskytu SNPs nalezených v našich vzorcích se SNPs s referenčním genomem *Erinaceus europeaus*, provedené pomocí Variant Effect Predictor (www.ensembl.org) je uvedeno v Tab. 11 a 12, grafy na Obr. 9 a 10 zobrazují průměrné hodnoty pro jednotlivé druhy.

Tab. 11. Celková anotace SNPs analyzovaných jedinců k referenčnímu genomu *E. europaeus*. Hodnoty ukazují, kolik procent zkoumaných SNPs patřilo k daným částem genomu. Pro jednoduchost byly ponechány názvy v anglickém jazyce.

ID	druh	intergenic variant [%]	intron variant [%]	downstream gene variant [%]	upstream gene variant [%]	synonymous variant [%]	missense variant [%]	non coding transcript variant [%]	splice region variant [%]	non coding transcript exon variant [%]
SK32	*E. roumanicus*	72	20	4	4	0	0	0	0	0
112	*E. roumanicus*	72	20	4	4	0	0	0	0	0
183	*E. roumanicus*	12	82	3	2	0	0	0	0	0
209	*E. roumanicus*	76	16	4	4	0	0	0	0	0
SK24	*E. roumanicus*	77	10	6	6	0	0	0	0	0
SK27	hybrid	72	20	4	4	0	0	0	0	0
122	*E. europaeus*	73	20	4	4	0	0	0	0	0
453	*E. europaeus*	73	20	4	4	0	0	0	0	0
211	*E. europaeus*	63	13	9	13	0	0	0	0	0
197	*E. europaeus*	74	19	3	3	0	0	0	0	0

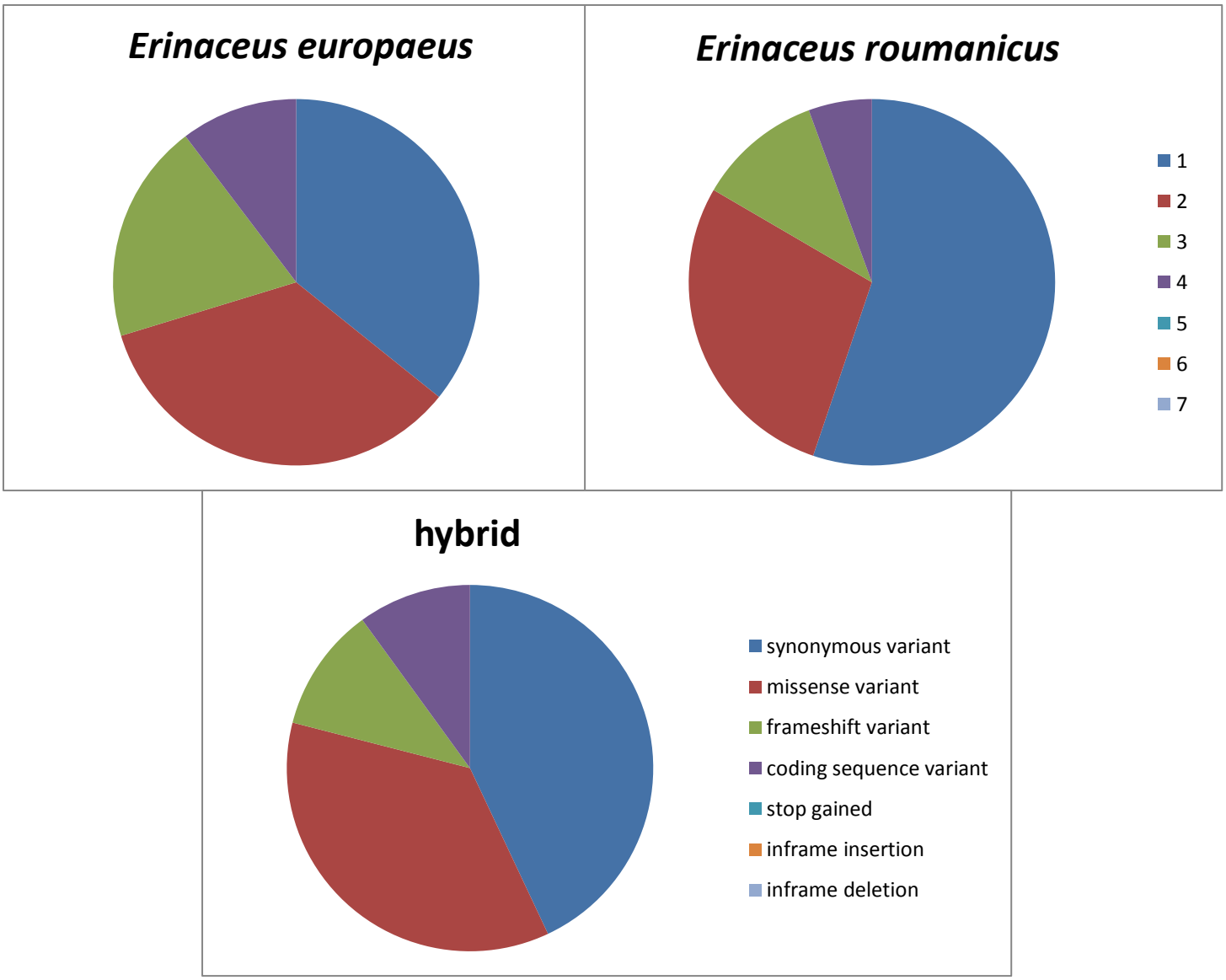
Tab. 12. Anotace SNPs zkoumaných jedinců ke kódujícím oblastem genomu referenčního genomu *E. europaeus*. Hodnoty ukazují, kolik procent zkoumaných SNPs patřilo k daným částem kódujícího genomu. Pro jednoduchost byly ponechány názvy v anglickém jazyce.

ID	druh	synonymous variant [%]	missense variant [%]	frameshift variant [%]	Coding sequence variant [%]	stop gained [%]	inframe insertion [%]	inframe deletion [%]
SK32	*E. roumanicus*	53	33	9	5	0	0	0
112	*E. roumanicus*	56	32	7	5	0	0	0
183	*E. roumanicus*	33	22	33	11	0	0	0
209	*E. roumanicus*	54	34	6	7	0	0	0
SK24	*E. roumanicus*	80	20	0	0	0	0	0
SK27	hybrid	43	36	11	10	0	0	0
122	*E. europaeus*	43	25	21	10	0	0	0
453	*E. europaeus*	33	35	19	12	0	0	0
211	*E. europaeus*	42	33	21	4	0	0	0
197	*E. europaeus*	24	44	16	15	0	0	0



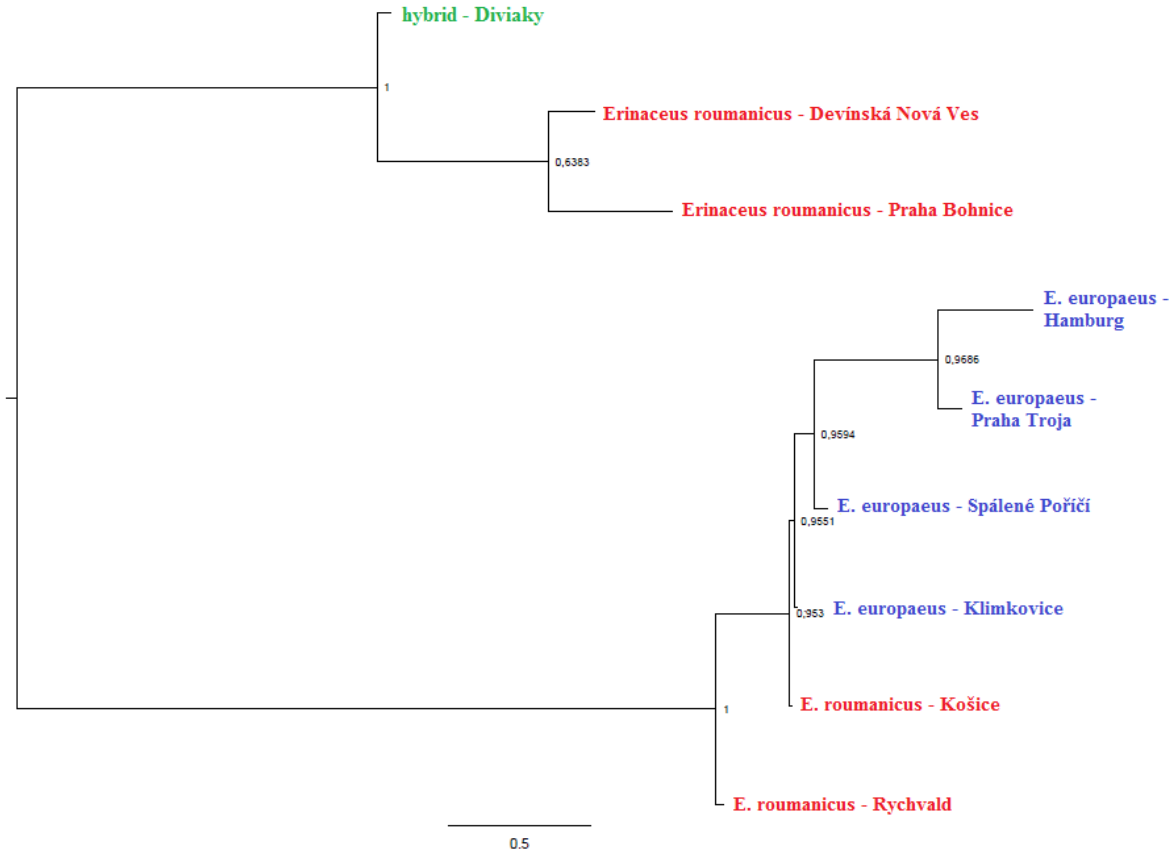
Obr. 11. Anotace SNPs navržených pro jednotlivé vzorky k referenčnímu genomu *E. europaeus*. Grafy zobrazují průměrné hodnoty pro *E. europaeus*, *E. roumanicus* a předpokládaného hybrida.

I přes mezidruhové rozdíly, byly SNPs nejčastěji v intergenických částech genomu, tedy v místech mezi známými geny a v intronech. Hybridní jedinec má podobné zastoupení jako druh *E. europaeus*.



Obr. 12. Anotace SNPs navržených pro jednotlivé vzorky k referenčnímu genomu *E. europaeus* pro kódující oblast DNA. Grafy zobrazují průměrné hodnoty pro *E. europaeus*, *E. roumanicus* a předpokládaného hybrida.

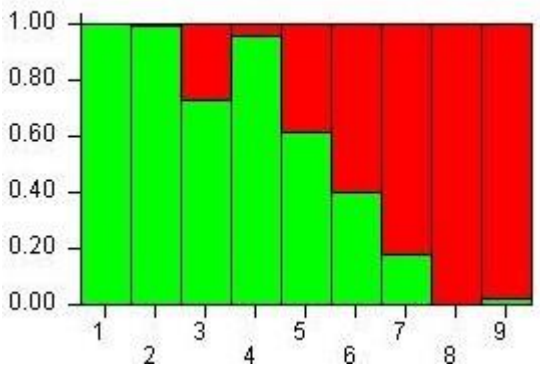
V kódujících oblastech byly u všech druhů nejčastěji zaznamenány synonymní mutace, tedy se stejnou výslednou aminokyselinou. Druhou nejčastější záměnou byly missence, tedy se změněnou aminokyselinou. U žádného druhu nebyly zjištěny mutace, které by vedly ke vzniku stop kodonu. Patrné jsou výrazné mezidruhové rozdíly.



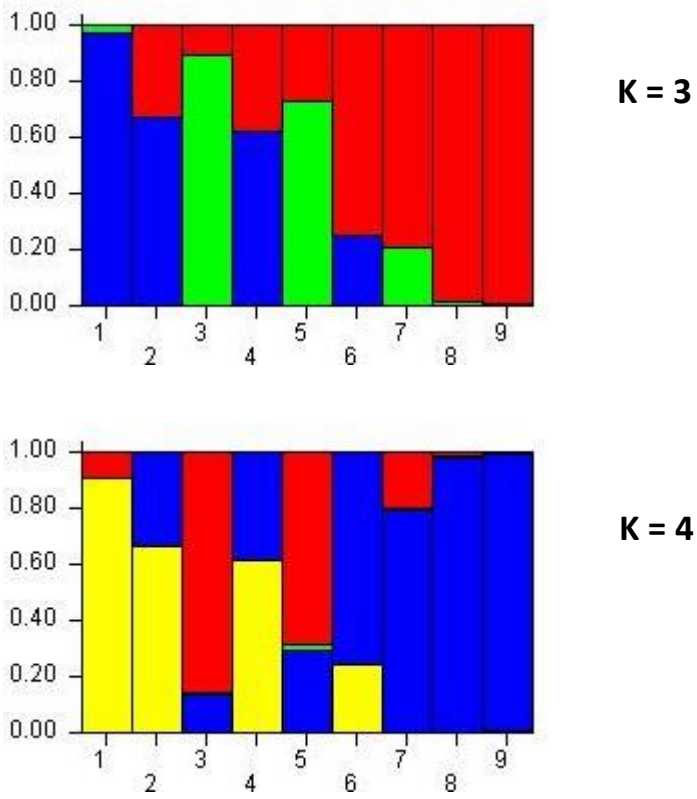
Obr. 13. Fylogenetický strom vytvořený v programu MrBayes (Ronquist *et al.* 2003). Bayesiánskou klastrovací analýzou. Zelenou barvou je označený hybrid, červenou *Erinaceus roumanicus* a modrou barvou pak *Erinaceus europaeus*.

SNPs navržené pro každý vzorek byly mezi sebou porovnány a bylo vytipováno 13268 SNPs, v jejichž přítomnosti se jednoznačně liší vzorky přiřazené k druhu *E. europaeus* a hybrid. Pro vzorky druhu *E. roumanicus* a hybrida byl počet rozdílných SNPs 7985.

Vzorek 183 (*E. roumanicus*) byl kvůli špatné kvalitě sekvence z dalších analýz vyřazen. Fylogenetický strom vytvořený pomocí programu MrBayes pro zbývajících 9 vzorků je znázorněn na Obr. 13. V této analýze hybrid klastroval s dvěma jedinci druhu *E. roumanicus*. Další dva jedinci druhu *E. roumanicus* ovšem klastrovali s *E. europaeus*.



K = 2



Obr. 14. Grafy zobrazující posteriorní pravděpodobnosti zařazení jedince do populace pro K = 2, K = 3 a K = 4, vytvořené pomocí programu Structure 2.3.4. Jednotlivé sloupce znázorňují hodnotu q, která přiřazuje jedince do dané subpopulace. Čísla sloupců znázorňují jednotlivé jedince v tomto pořadí: 1 - sk32 (*E. roumanicus*), 2 - 112 (*E. roumanicus*), 3 - 453 (*E. europaeus*), 4 - sk27 (hybrid), 5 - 122 (*E. europaeus*), 6 - 209 (*E. roumanicus*), 7 - 197 (*E. europaeus*), 8 - 211 (*E. europaeus*), 9 - sk24 (*E. roumanicus*).

Rozdělení datasetu v programu Structure znázorňuje Obr. 14. Nejvyšší podpora byla zjištěna pro K = 2, tedy rozdělení datasetu do dvou subpopulací. Hodnoty podpory pro jednotlivá K jsou uvedeny v Tab. 13.

Tab. 13. Hodnoty podpory pro jednotlivé klastry (K) v programu Structure.

	K = 2	K = 3	K = 4
Estimated Ln Prob of Data	-6298875.0	-12413720.9	-81366.7

4 Diskuze

Většina námi nalezených SNPs po srovnání s referenčním genomem *E. europaeus* v databázi Ensembl patřila do intergenických oblastí a do oblastí intronů, tedy do nekódujících části genomu - průměrně 91,4% sekvencí pro *E. roumanicus* a 88,75% pro *E. europaeus*. Při srovnávání kódujících oblastí tvořily převážnou část sekvence se synonymní záměnou, tedy ve výsledku obsahující stejnou aminokyselinu - pro *E. roumanicus* to bylo 55,2% sekvencí, pro *E. europaeus* 35,5%. Záměny s rozdílnou aminokyselinou tvořily u *E. roumanicus* 28,2%. U *E. europaeus* to bylo 34,25% a rozdíl mezi synonymními a nesynonymními záměnami byl tak minimální. Pro účely populační genetiky, která si klade za cíl zjišťovat detaily o populační struktuře volně žijících druhů je využití neutrálních markerů, tedy takových oblastí, které nepodléhají selekci, výhodou. Nalezené SNPs však mohou být ve vazbě s některými geny a ve skutečnosti se tedy nemusí chovat jako selekčně neutrální. Před dalšími analýzami je nutné tuto skutečnost otestovat.

Byla zjištěna korelace mezi naměřenou koncentrací DNA a výsledným počtem nalezených SNPs. Tato skutečnost není překvapivá, avšak tato korelace nebyla prokázána na přesnějším analyzátoru Qubit. Zjištěná korelace tak je zřejmě velmi slabá a výsledná kvalita a pokrytí genomu závisí i na dalších proměnných než pouze na koncentraci DNA, například na čistotě DNA, integritě apod. Integrita byla posuzována pomocí gelové elektroforézy. Pro sekvenaci byly vybrány pouze vzorky, u kterých se vyskytovali fragmenty DNA delší než přibližně 500bp.

Při použití vstupního souboru ve formátu obsahujícím pouze shody a neshody oproti referenčnímu genomu byl zjištěn výrazný nesoulad s předpoklady výstupů těchto analýz. Druhové určení na základě kontrolního úseku mitochondriální DNA, které bylo ověřeno v předchozích studiích (Bolfikova & Hulva 2012), se neshodovalo s výsledky fylogenetické analýzy provedené Bayesiánskou klastrovací metodou. Posteriorní podpory pro jednotlivé štěpení byly v hodnotách, které se obecně považují za dostatečné, avšak výsledek byl v rozporu s druhovým statusem daného jedince. Dva jedinci *E. roumanicus* klastrovali k bázi *E. europaeus* a hybridní jedinec klastroval k bázi druhu *E. roumanicus*. Předchozí analýza mikrosatelitových markerů naznačila větší zastoupení genomu *E. europaeus* u tohoto jedince. Mitochondriální DNA byla taktéž druhu *E. europaeus*. Hybrid použitý v našem datasetu pochází z obce Diviaky na Slovensku, která se nachází na samém okraji sympatrické zóny druhů *E. europaeus* a *E. roumanicus* ve střední Evropě. Předpokladem předchozích

analýz bylo, že se jedná o potomka samice, která na kraji areálu nenalezla samce svého druhu a spářila se s *E. roumanicus*. Vyšší afiliaci ke genomu *E. europaeus* naznačovala i anotace nalezených SNP k referenčnímu genomu *E. europaeus*. K interpretaci těchto zjištění bude potřeba dalších analýz. První možností je výskyt nějakého artefaktu použité metody. Je možné, že ztráta mnoha informací při převodu z nukleotidů na binární kód způsobila nejednoznačnosti v Bayesiánské analýze. Tato metoda se tedy jeví jako nevhodná pro fylogenetická srovnávání a to i s pokročilým metodickým rámcem, jako je Bayesovská analýza. Další možností je efekt použitého vzorkování, které zahrnuje vzorky z jedinců uhynulých na silnicích, vzorky kůže atd., které mohou obsahovat mikrobiální DNA. Tento vliv by měl být odstraněn přiřazením sekvencí referenčnímu genomu cílového druhu, tuto skutečnost ale bude třeba ověřit zpracováním velkého množství vzorků a testováním vlivu použité tkáně (např. sval vs. kůže).

Další možností je ovšem komplikovanější populačně genomická architektura na zkoumaném území než bylo předpokládáno na základě analýz klasických genetických markerů. Například jednotlivé interglaciální periody mohly mít za následek vznik kontaktní zóny s jinými geografickými i biologickými parametry a mohlo docházet k následným vlnám introgrese před vznikem dnes poměrně neprostupné reprodukčně izolační bariéry (Bolfikova & Hulva 2012). K ověření těchto skutečností bude potřeba zpracovat vzorky s větším geografickým pokrytím zkoumaného areálu.

Pro zjištění detailnější genetické struktury byl využit program Structure, taktéž fungující na principu Bayesiánké klastrovací analýzy. Přestože mikrosatelitové markery v předchozích pracích jednoznačně odlišily druhy (Bolfikova & Hulva 2012), zde nedošlo k rozdělení mezidruhovému a celkově se struktura se zvyšujícím se počtem klastrů nedala korelovat s jasným biologickým významem. Zde je opět obtížné rozlišit, zda tento fakt má metodickou nebo objektivní příčinu. První možností je vliv nevhodně zvoleného binárního vstupního souboru, kde došlo ke ztrátě informací. Další možností je opět komplikovaná populačně genomická architektura. Naznačená populační struktura při nejlépe podpořeném K = 2 naznačuje určité kontinuum, a tedy možnost existence hybridního roje. V předchozích analýzách byly hypotézy o evoluční minulosti skupiny formulovány na základě použití devíti mikrosatelitových lokusů. Genom ježka má však 48 chromozomů (Geisler & Gropp 1967).

To znamená, že bez přítomnosti genové vazby mezi zkoumanými lokusy bylo pokryto méně než 20% z chromozomové sady zkoumaného druhu, a to pouze jedním lokusem. Mitochondriální DNA je zase haploidní nerekombinující molekula a vzhledem k mnoha svým specifikům má značné limity při interpretačních přechodech od genových k druhovým evolučním historiím (Avise 1992). Vzhledem ke komplikovaným genetickým důsledkům jednotlivých evolučních procesů včetně introgrese, rekombinace apod. Je jasné, že práce založené na klasických markerech představují hrubou aproximaci. Např. koncept porézního genomu a genomických ostrovů formalizuje pozorování že introgrese jednotlivých lokusů závisí na jejich vzdálenosti od oblastí výskytu genů zodpovědných za ekologickou a behaviorální izolaci (Turner et al. 2005). Bloky přispívající k izolaci se spíše nacházejí v oblastech se sníženou úrovní rekombinace, jako jsou chromozomální inverse a centromerické oblasti (Navarro & Barton 2003). Alternativní hypotézy založené na matematických a simulačních modelech (Feder & Nosil 2010) i empirických studiích (Parchman *et al.* 2013) naopak předpokládají, že lokusy zodpovědné za adaptivní divergence a reprodukční izolaci jsou rozptýlené v genomu. Tyto diskrepance naznačují, že způsoby genomové diferenciace v alopatrii a genomických interakcí v sympatrii jsou taxonově specifické. V každém případě při malém počtu genetických markerů tedy může být výsledek značně ovlivněn náhodou (lokalizací zkoumaných lokusů v genomu).

Genomické markery s o několik řádů větším pokrytím genomu mohou tedy přinést mnohem adekvátnější znázornění reality.

Pro testování těchto hypotéz bude potřeba kromě dalších dat i velký objem bioinformatické práce. V dalším výzkumu plánujeme použít software Stacks (Catchen *et al.* 2013), fungující v operačním systému Linux, který k identifikaci polymorfismů v sekvenci využívá statistického modelu maximální pravděpodobnosti (maximum likelihood).

5 Přehled literatury

Aronesty E (2011) ea-utils : Command-line tools for processing biological sequencing data. *Expression Analysis*, Durham.
http://code.google.com/p/ea-utils.

Aulagnier S, Haffner P, Mitchell-Jones AJ, Moutou F, Zima J (2009) Mammals of Europe, North Africa and the Middle East. *A & C Black Publisher Ltd*, London, 42.

Avise JC (1992) Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *Oikos*, **63**, 62-76.

Bannikova AA, Lebedev VS, Abramov AV, Rozhnov VV (2014) Contrasting evolutionary history of hedgehogs and gymnures (Mammalia: Erinaceomorpha) as inferred from a multigene study. *Biological Journal of the Linnean Society* **112**, 499-519.

Bell RC, Drewes RC, Zamudio KR (2015) Reed frog diversification in the Gulf of Guinea: Overseas dispersal, the progression rule, and in situ speciation. *Evolution* **69**, 904-915.

Bentley DR, Balasubramanian S, Swerdlow HP*, et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59.

Berggren KT, Ellegren H, Hewitt GM, Seddon JM (2005) Understanding the phylogeographic patterns of European hedgehogs, Erinaceus concolor and E-europaeus using the MHC. *Heredity* **95**, 84-90.

Boehm JT, Waldman J, Robinson JD, Hickerson MJ (2015) Population Genomics Reveals Seahorses (Hippocampus erectus) of the Western Mid-Atlantic Coast to Be Residents Rather than Vagrants. *Plos One* **10**.

Bogdanov AS, Bannikova AA, Pirusskii YM, Formozov NA (2009) The first genetic evidence of hybridization between West European and Northern white-breasted hedgehogs (Erinaceus europaeus and E. roumanicus) in Moscow region. *Biology Bulletin* **36**, 647-651.

Bolfikova B, Hulva P (2012) Microevolution of sympatry: landscape genetics of hedgehogs Erinaceus europaeus and E. roumanicus in Central Europe. *Heredity* **108**, 248-255.

Bolfikova B, Konecny A, Pfaffle M, Skuballa J, Hulva P (2013) Population biology of establishment in New Zealand hedgehogs inferred from genetic and historical data: conflict or compromise? *Molecular Ecology* **22**, 3709-3720.

Brockie RE (1975) Distribution and abundance of the hedgehog (*Erinaceus europaeus*) in New Zealand 1869, 1973. *New Zealand Journal of Zoology* **2**, *445–462*.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Colosimo PF, Hosemann KE, Balabhadra S*, et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928-1933.

Černá Bolfíková B (2013) Evoluční historie ježků rodu *Erinaceus*. Disertační práce. *Univerzita Karlova v Praze,* Praha.

Davey JL, Blaxter MW (2010) RADSeq: next-generation population genetics. *Briefings in Functional Genomics* **9**, 416-423.

Eliášová K (2014) Vnitřní struktura balkánského refugia na modelu *Erinaceus roumanicus.* Diplomová práce. *Univerzita Karlova v Praze*, Praha.

Emerson KJ, Merz CR, Catchen JM*, et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16196-16200.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186-194.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.

Field D, Tiwari B, Booth T*, et al.* (2006) Open software for biologists: from famine to feast. *Nature Biotechnology* **24**, 801-803.

Galik A, Mohandesan E, Forstenpointner G*, et al.* (2015) A Sunken Ship of the Desert at the River Danube in Tulln, Austria. *Plos One* **10**.

Gall A, Ferns B, Morris C*, et al.* (2012) Universal Amplification, Next-Generation Sequencing, and Assembly of HIV-1 Genomes. *Journal of Clinical Microbiology* **50**, 3838-3844.

Geisler M, Gropp A (1967) Chromosome Polymorphism in the European Hedgehog Erinaceus europaeus (Insectivora). *Nature* **214**, 396-397.

GISD (2015) Global invasive species database http://www.issg.org/database/welcome/

Herter K (1965) Hedgehogs, a comprehensive study. *Phoenix house*.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-913.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68**, 87-112.

Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *Plos Genetics* **6**.

Illumina Inc. (2010) Illumina Sequencing Technology. Pub. No. 770-2007-002 http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina Inc. (2011) Quality scores for Next-generation Sequencing: Assessing sequencing accuracy using Phred quality scoring. Pub. No. 770-2011-030

http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Illumina Inc. (2014) Sequencing power for every scale. Pub. No. 770-2011-030 http://www.illumina.com/content/dam/illuminamarketing/documents/products/brochures/brochure_sequencing_systems_portfolio.pdf

IUCN (2015) The IUCN Red List of Threatened species. Version 2015.2 http://www.iucnredlist.org/

Jackson DB (2001) Experimental removal of introduced hedgehogs improves wader nest success in the Western Isles, Scotland. *Journal of Applied Ecology* **38**, 802-812.

Kovach AI, Walsh J, Ramsdell J, Thomas WK (2015) Development of diagnostic microsatellite markers from whole-genome sequences of Ammodramus sparrows for assessing admixture in a hybrid zone. *Ecology and Evolution* **5**, 2267-2283.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595.

Li H, Handsaker B, Wysoker A*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Lowe S, Browne M, Boudjelas S, De Poorter M (2000) 100 of the World's Worst Invasive Alien Species: A selection from the Global Invasive Species Database. The Invasive Species Specialist Group (ISSG) a specialist group of the Species Survival Commission (SSC) of the World Conservation Union (IUCN).

Lozier JD (2014) Revisiting comparisons of genetic diversity in stable and declining species: assessing genome-wide polymorphism in North American bumble bees using RAD sequencing. *Molecular Ecology* **23**, 788-801.

Lozier JD, Strange JP, Stewart IJ, Cameron SA (2011) Patterns of range-wide genetic variation in six North American bumble bee (Apidae: Bombus) species. *Molecular Ecology* **20**, 4870-4888.

Madsen O, Scally M, Douady CJ*, et al.* (2001) Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610-614.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133-141.

Morgan JE, Carr IM, Sheridan E*, et al.* (2010) Genetic Diagnosis of Familial Breast Cancer Using Clonal Sequencing. *Human Mutation* **31**, 484-491.

Murphy WJ, Eizirik E, Johnson WE*, et al.* (2001a) Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614-618.

Murphy WJ, Eizirik E, O'Brien SJ*, et al.* (2001b) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348-2351.

Navarro A, Barton NH (2003) Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes. *Science*, **300**, 321–324.

Parchman TL, Gompert Z, Braun MJ*, et al.* (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology* **22**, 3304-3317.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *Plos One* **7**.

Pfaeffle M, Bolfikova BC, Hulva P, Petney T (2014) Different Parasite Faunas in Sympatric Populations of Sister Hedgehog Species in a Secondary Contact Zone. *Plos One* **9**.

Poduschka W & Poduschka C (1983) Kreuzungsversuche an mitteleuropäischen Igeln. *Säugetierk Mitt*, **31**, 1–12.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rackham O, Moody J (1996) The Making of the Cretan Landscape, *Manchester University Press*, Manchester, 47.

Reeve N (1994) Hedgehogs. *T & AD Poyser (Natural History)*, London, 17.

Reis-Filho JS (2009) Next-generation sequencing. *Breast Cancer Research* **11**.

Robinson JT, Thorvaldsdottir H, Winckler W*, et al.* (2011) Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26.

Roca AL, Bar-Gal GK, Eizirik E*, et al.* (2004) Mesozoic origin for West Indian insectivores. *Nature* **429**, 649-651.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574. doi: 10.1093/bioinformatics/btg180

Ruiz E, Mohandesan E, Fitak RR, Burger PA (2015) Diagnostic single nucleotide polymorphism markers to identify hybridization between dromedary and Bactrian camels. *Conservation Genetics Resources* **7**, 329-332.

Santucci F, Emerson BC, Hewitt GM (1998) Mitochondrial DNA phylogeography of European hedgehogs. *Molecular Ecology* **7**, 1163-1172.

Seddon JM, Santucci F, Reeve N, Hewitt GM (2002) Caucasus Mountains divide postulated postglacial colonization routes in the white-breasted hedgehog, Erinaceus concolor. *Journal of Evolutionary Biology* **15**, 463-467.

Seddon JM, Santucci F, Reeve NJ, Hewitt GM (2001) DNA footprints of European hedgehogs, Erinaceus europaeus and E-concolor. Pleistocene refugia, postglacial expansion and colonization routes. *Molecular Ecology* **10**, 2187-2198.

Suchentrunk F, Haiden A, Hartl GB (1998) On biochemical genetic variability and divergence of the two Hedgehog species Erinaceus europaeus and E-concolor in central Europe. *Zeitschrift Fur Saugetierkunde-International Journal of Mammalian Biology* **63**, 257-265.

Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178-192.

Turner TL, Hahn MW, Nuzhdin (2005) Genomic islands of speciation in Anopheles gambiae. *PLOS Biology,* **3**, 285.

Vitousek PM, Dantonio CM, Loope LL, Westbrooks R (1996) Biological invasions as global environmental change. *American Scientist* **84**, 468-478.

Wilgenbusch JC, Warren DL, Swofford DL (2004) AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.

Woiss CH (2007) StatSoft, Inc., Tulsa, OK.: STATISTICA, version 8. *Asta-Advances in Statistical Analysis* **91**, 339-341.

Wu J, Li L-T, Li M*, et al.* (2014) High-density genetic linkage map construction and identification of fruit-related QTLs in pear using SNP and SSR markers. *Journal of Experimental Botany* **65**, 5771-5781.

6 Příloha

Appendix 1. **Výstup z aplikace RADtag counter from GenePool**, pomocí které byla vybrána nejvhodnější restriktáza (SbfI) pro štěpení ježčího genomu.

RADtag counter from GenePool, Edinburgh

To use this counter:

1	Enter the GC content of your target genome here:	**0,42**	proportion GC
2	Enter the size in megabases of your genome here:	**2708**	megabases genome
3	Enter the fold coverage of RADtags you require here:	**30**	fold coverage
4	Enter the per-pool plexity you plan to use here:	**25**	plexity
5	Enter number of million reads per lane		
	(please contact the GenePool for throughput currently	**100**	million reads per lane
	achieved on the GAIIx and HiSeq platforms)		

Overhang	**TGCA**			**GGCC**			**AATT**			**GATC**				**TCA**
Enzyme	**SbfI**	**PstI**	**NsiI**	**NotI**	**EaeI**	**EagI**	**EcoRI**	**ApoI**	**MfeI**	**BamHI**	**BclI**	**BglII**	**BstYI**	**BbvCI**
Site	CCTGCA*GG	CTGCA*G	ATGCA*T	GC*GGCCGC	Y*GGCCR	C*GGCCG	G*AATTC	R*AATTY	C*AATTG	G*GATCC	T*GATCA	A*GATCT	R*GATCY	CC*TCAGC
Site frequency	7,21E-06	0,000164	0,000312	3,78E-06	0,000486	8,58E-05	0,000312	0,0017682	0,000312	0,000164	0,000312	0,000312	0,0009272	3,43E-05
Sites/Mb	7	164	312	4	486	86	312	1768	312	164	312	312	927	34
Number of sites in genome	**19533**	**442916**	**844655**	**10242**	**1316636**	**232255**	**844655**	**4788292**	**844655**	**442916**	**844655**	**844655**	**2510864**	**93012**
Number of tags	39065	885833	1689310	20485	2633273	464509	1689310	9576585	1689310	885833	1689310	1689310	5021729	93012
Num sequences for coverage	1171957	26574988	50679286	614546	78998182	13935279	50679286	287297542	50679286	26574988	50679286	50679286	150651862	2790374
Million sequences per pool	29,3	664,4	1267,0	15,4	1975,0	348,4	1267,0	7182,4	1267,0	664,4	1267,0	1267,0	3766,3	69,8
does your pool fit in one lane?	**YES**	**NO**	**NO**	**YES**	**NO**	**NO**	**NO**	**NO**	**NO**	**NO**	**NO**	**NO**	**NO**	**YES**

Appendix 2. **Sequenced RAD Markers for Rapid SNP Discovery and Genetic Mapping**

Paul D. Etter (modified by L. Choleva for *SbfI* rextriction enzyme)

2. Materials

2.1. DNA extraction and RNase A treatment

1. DNeasy Blood & Tissue Kit (Qiagen).
2. RNaseA (Qiagen).

2.2. Restriction endonuclease digestion

1. Restriction enzyme (NEB): ***Sbf*I-HF**.
2. Clean, intact high-quality genomic DNA: 25 ng/µl.

2.3. P1 Adapter ligation

1. NEB Buffer 2.
2. rATP (Promega): 100 m*M*.
3. P1 Adapter: 100 n*M.* A modified Solexa© adapter (2006 Illumina, Inc., all rights reserved). Prepare 100nM stocks of P1 adapters in 1X Annealing Buffer (AB, *see* **Note 4**).

Below, example barcoded *Sbf*I P1 adapter sequences. "P" denotes a phosphate group and "x" refers to barcode nucleotides.

P1 top:

5´-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxTGC*A-3´

P1 bottom:

5´-Phos-xxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T-3´

The oligos are given in the excel file "Oligo Table RAD_P1_adapters_Lukas Choleva.xls"

4. Concentrated T4 DNA Ligase (NEB): 2,000,000 U/ml.

2.4. Purification steps

1. QIAquick or MinElute PCR Purification Kit (Qiagen).

2.5. DNA shearing

1. Bioruptor, nebulizer or Branson sonicator 450.

2.6. Size selection/agarose gel extraction

1. Agarose (Sigma)
2. 5X TBE: 0.45 *M* Tris-Borate, 0.01 *M* EDTA, pH 8.3.
3. 6X Orange Loading Dye Solution (Fermentas).
4. GeneRuler 100 bp DNA Ladder Plus (Fermentas).
5. Razor blades.
6. MinElute Gel Purification Kit (Qiagen).

2.7. Perform end repair

1. Quick Blunting Kit (NEB).

2.8. 3´-dA overhang addition

1. NEB Buffer 2.
2. dATP (Fermentas): 10 m*M.*
3. Klenow Fragment (3´ to 5´ exo$^-$, NEB): 5,000 U/ml.

2.9. P2 Adapter ligation

1. NEB Buffer 2.
2. rATP: 100 m*M.*

3. P2 Adapter: 10 µ*M.* A modified Solexa© adapter (2006 Illumina, Inc., all rights reserved). Prepare 10 µ*M* double-stranded adapter in 1X AB (*see* **Note 4**). Asterisk denotes a phosphorothioate bond introduced to confer nuclease resistance to the double-stranded oligo ***(14)***.

Paired End

P2 top:

5´- /5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA-3´

P2 bottom:

5´CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T -3´

The oligos are also given in the excel file "Oligo Table RAD_P2_adapters_Lukas Choleva.xls"

4. Concentrated T4 DNA Ligase.

2.10. RAD tag Amplification/Enrichment

1. Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB).

2. Modified Solexa© Amplification primer mix (2006 Illumina, Inc., all rights reserved): 10 µ*M.*

P1-forward primer: 5´- AATGATACGGCGACCACCG*A -3´

P2-reverse primer: 5´- CAAGCAGAAGACGGCATACG*A -3´

The oligos are also given in the excel file "Oligo Table RAD_P2_adapters_Lukas Choleva.xls"

3. Methods (GENERAL)

The protocol described below, outlined in **Fig. 1,** prepares RAD tag libraries for high-throughput Illumina sequencing (*see* **Note 1**). In short, genomic DNA is digested with a restriction enzyme and an adapter (P1) is ligated to the fragment's compatible ends (**Fig. 1A**). This adapter contains forward amplification and Illumina sequencing primer sites, as

well as a nucleotide barcode 4 or 5 bp long for sample identification. To reduce erroneous sample assignment due to sequencing error, all barcodes differ by at least two nucleotides. The adapter-ligated fragments are subsequently pooled, randomly sheared, and size-selected (**Fig. 1B**). DNA is then ligated to a second adapter (P2), a Y adapter ***(13)*** that has divergent ends (**Fig. 1C**). The reverse amplification primer is unable to bind to P2 unless the complementary sequence is filled in during the first round of forward elongation originating from the P1 amplification primer. The structure of this adapter ensures that only P1 adapter-ligated RAD tags will be amplified during the final PCR amplification step (**Fig. 1D**). The protocol for mapping of the lateral plate locus in stickleback using *EcoR*I RAD markers used in Baird et al., 2008 ***(12)*** is described here in detail as an example of the multiplexing approach. For bulk-segregant analysis pooled samples are combined prior to digestion and treated as a single library with one barcode.

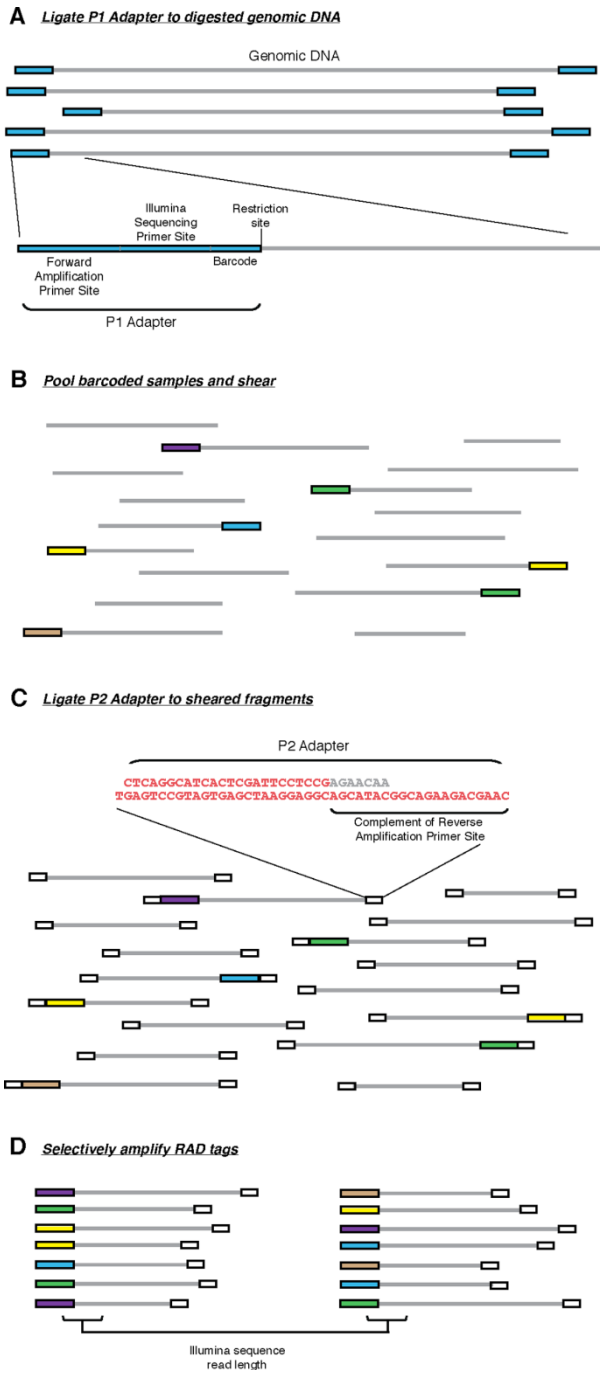


Fig. 1. RAD tag library generation. (**A**) Genomic DNA is digested with a restriction enzyme and a barcoded P1 adapter is ligated to the fragments. The P1 adapter contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode (colored boxes represent P1 adapters with different barcodes). (**B**) Adapter-ligated fragments are combined (if multiplexing), sheared and (**C**) ligated to a second adapter (P2, white boxes). The P2 adapter is a divergent "Y" adapter, containing the reverse complement of the P2 reverse amplification primer site, preventing amplification of genomic fragments

3.1. DNA extraction and RNase A treatment

1. **IGNORE** We recommend extracting genomic DNA samples using the DNeasy Blood & Tissue Kit (Qiagen) or a similar product that produces very pure, high molecular weight, RNA-free DNA. High-quality DNA is required for optimal restriction endonuclease digestion and is of utmost importance for the overall success of the protocol. Follow the manufacturers instructions for extraction from your tissue type. Be sure to treat samples with RNase A following manufacturer's instructions to remove residual RNA. Quantify the DNA using a fluorimeter to get the most accurate concentration readings (*see* **Note 3**). The optimal concentration after elution is 25 ng/µl or greater.

3.1.5 Anneal adapters

Single-stranded oligos need to be annealed with their appropriate partner before ligation. We provide sequences for 48 uniquely barcoded adapter P1 oligo pairs (oligos P1-FOR and P1-REV) and common adapter P2 (oligos P2-FOR and P2-REV), see **OLIGO TABLE RAD_P1_adapters.**

1. To create Adapter P1, combine each oligo P1-FOR with its complementary oligo P1-REV and in a 1:1 ratio in working strength annealing buffer (final buffer concentration 1x) for a total annealed adapter concentration of 40uM (for example, if purchased oligos are resuspended to an initial concentration of 200uM, use 20ul oligo P1-FOR, 20ul oligo P1-REV, 10ul 10x annealing buffer and 50ul nuclease-free water). Do the same for oligos P2-FOR and P2-REV to create the common adapter P2.

2. In a thermocyler, incubate at 97.5°C for 2.5 minutes, and then cool at a rate of not greater than 3°C per minute until the solution reaches a temperature of 21°C. Hold at 4°C.

3. Prepare final working strength concentrations of annealed adapters from this annealed stock (the appropriate working stock dilution for your experiment can be determined from our ligation molarity calculator). For convenience, it is possible to store the adapters at 4°C while in active use.

3.2. Restriction endonuclease digestion

1. Digest 0.1-1µg genomic DNA for each individual sample (**for 60 min at 37°C in a 50 µl reaction volume containing 5.0 µl 10× Buffer 4 and 10 units (U) *Sbf*I-HF (New England Biolabs [NEB])**, following the manufacturers instructions.

(alternatively, to ensure complete digestion, run digests for 3 hours at 37°C, holding at 4°C. Do not heat kill the enzymes, as this may skew base composition in the resulting fragment library. Before proceeding with step "Clean the double digest with AMPure XP beads", cool the reaction to room temperature. Alternatively, reactions can be stored at 4°C overnight).

2. Heat-inactivate the restriction enzyme following manufacturer's instructions. **(Heat-inactivate for 20 min at 65°C).** Allow reaction to cool slowly to ambient temperature (30-60 min). If the enzyme cannot be heat-inactivated, purify with a QIAquick column following manufacturer's instructions prior to ligation.

3.3. P1 Adapter ligation

1. This step in the protocol ligates barcoded, restriction-site overhang specific P1 adapters onto complementary compatible ends on the genomic DNA created in the previous step (*see* **Note 5**). 48 different barcoded P1 adapters are used to make ***Sbf*I-HF** RAD tag libraries for 48 individuals (*ignore* Note 7).

2. To each inactivated digest, add:

3.0 µl (2.5 µl – used by Baird et al. 2008) Barcoded SbfI-P1 Adapter (100 nM), a modified Illumina© adapter (2006 Illumina, Inc., all rights reserved; top oligo: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxTGC*A-3' [xxxxx = barcodes, * = phosphorothioate bond]; bottom oligo: 5'-Phos-xxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT*T-3'), added to each sample along with 0.6 µl rATP (100 m*M*, Promega), 1.0 µl 10× SbfI Buffer 4, 0.5 µl (1000 U; 2,000,000 U/ml) T4 DNA Ligase (high concentration, NEB), 4.9 µl H2O; 60.0 µl total volume; and incubate reaction at room temperature (RT) for 30 min.

(note: some authors add 750 pmol P1 adaptor per 1 ug)

3. NEB Buffer 4 is used in the ligation reactions in this protocol instead of ligase buffer because the salt it contains (50 m*M* NaCl) ensures the double-stranded adapters remain

annealed during the reactions (*see* **Note 4**). T4 DNA Ligase is active in all 4 NEB Buffers if supplemented with 1m*M* rATP. If the restriction buffer used for digestion does not contain at least 50 m*M* potassium or sodium ions, or if the endonuclease cannot be heat-inactivated and the reaction must be purified in a column prior to P1 ligation, add 6.0 µl NEB Buffer 4.

4. **IGNORE:** Reduce the amount of P1 used in the ligation reaction if starting with less than 1µg genomic DNA or if cutting with an enzyme that cuts less frequently than *EcoR*I (**yes, this is the case for SbfI, as this enzyme cuts more frequently than *EcoR*I . Therefore, 4.0 µl Barcoded SbfI-P1 Adapter is used instead of 5.0 µl, as typical when *EcoR*I enzyme is used**). It is critical to optimize the amount of P1 adapter added when a given restriction enzyme is used for the first time in an organism (*see* **Note 6**).

5. Heat-inactivate T4 DNA Ligase for 20 min at 65° C. Allow reaction to cool slowly to ambient temperature (30 min).

3.4. Sample multiplexing

1. This step allows multiple individually barcoded samples to be combined and processed as one to cut down on cost, work time, and differences in amplification efficiency that may arise between different library preparations when processing many at once.

2. Combine barcoded samples at desired ratio. Use a 100-300 µl aliquot containing 1-2 µg DNA total to complete the protocol and freeze the rest at -20° C.

(**IGNORE:** In Baird et al., 2008 ***(12)*** F_0 parent samples, as well as the F_2 pools used for bulk-segregant analysis, were combined at equal volumes to create one library (*see* **Fig. 2**, lanes 2, 3 & 5). *EcoR*I libraries containing barcoded samples from F_2 individuals sharing a given lateral plate phenotype were pooled and processed as separate libraries after P1 ligation (*see* **Note 7**)).

3.5. DNA shearing

1. Shear DNA samples to an average size of 500 bp to create a library of P1/restriction-site-ligated molecules with random variable ends for amplification. This step requires some optimization for different DNA concentrations and each time a different restriction endonuclease is used. The following protocol has been optimized to shear Stickleback DNA

digested with (either *EcoR*I) or ***Sbf*I** using the Bioruptor and is a good starting point for any study. The goal is to create sheared product that is predominantly smaller than 1 kb in size (*see* **Fig. 2**).

2. Dilute ligation reaction to 100 µl in water (or take 100-300 µl aliquot from multiplexed samples) and shear in Bioruptor 10 times for 30 sec on high following manufacturer's instructions.

(note: some authors shear using five 30 s on-and-off cycles)

3. Clean up sheared DNA sample(s) using a MinElute column following manufacturer's instructions. This purification is performed in order to remove the ligase and restriction enzymes, and to concentrate the DNA so that the entire sample can be loaded in a single lane on an agarose gel. Elute in 20 µl EB.

3.6. Size selection/agarose gel extraction

1. This step in the protocol removes free un-ligated or concatomerized P1 adapters and restricts the size range of tags to that which can be sequenced efficiently on an Illumina Genome Analyzer flow cell. Run the entire sheared sample in 1X Orange Loading Dye on a 1.25% agarose, 0.5X TBE gel for 45 min at 100 V, next to 2.0 µl GeneRuler 100 bp DNA Ladder Plus for size reference (*see* **Fig. 2**, **Note 8**).

2. Being careful to exclude any free P1 adapters and P1 dimers running at ~130 bp and below, use a fresh razor blade to cut a slice of the gel spanning 300-500 (-700) bp. Extract DNA using MinElute Gel Purification Kit following manufacturer's instructions with the following modification: to improve representation of A + T-rich sequences, melt agarose gel slices in the supplied buffer at room temperature (18-22° C) with agitation for 30 min ***(14)***. **Elute in 19 µl EB into eppendorf tube containing 2.5 µl 10X Blunting Buffer from Quick Blunting Kit used in the following step.**

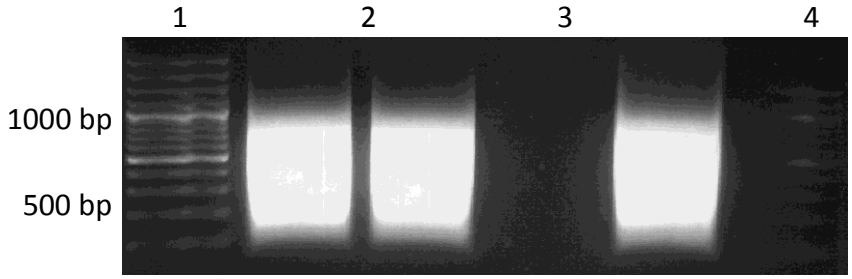


Fig. 2. Three barcoded and multiplexed RAD tag libraries. **2**, **3** & **5** each contain two DNA samples that were restriction digested, ligated to barcoded P1 adapters, combined, sheared, purified and then loaded on an agarose gel. **2** - F_0 parental DNA samples cut with *Sbf*I. **3** - F_2 pools cut with *Sbf*I. **4** - blank. **5** - F_0 parental DNA samples cut with *EcoR*I. Libraries

3.7. Perform end repair (The Quick Blunting Kit (NEB) was used to polish the ends of the DNA.)

1. The Quick Blunting Kit protocol converts 5´ or 3´ overhangs, created by shearing, into phosphorylated blunt ends using T4 DNA Polymerase and T4 Polynucleotide Kinase.

2. To the eluate from the previous step, add: 2.5 µl dNTP mix (1m*M*), 1.0 µl Blunt Enzyme Mix. Incubate at RT for 30 min.

3. Purify with QIAquick column. Elute in 43 µl EB into eppendorf tube containing 5.0 µl 10X NEB Buffer 2.

3.8. 3´-dA overhang addition

1. This step in the protocol adds an 'A' base to the 3´ ends of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow Fragment (3´ to 5´ exo$^-$). This prepares the DNA fragments for ligation to the P2 adapter, which possesses a single 'T' base overhang at the 3´ end of its bottom strand.

2. To the eluate from the previous step, add: 1.0 µl dATP (10m*M*), 3.0 µl Klenow (exo$^-$). Incubate at 37º C for 30 min. Allow reaction to cool slowly to ambient temperature (15 min).

3. Purify with QIAquick column. Elute in 45 µl EB into eppendorf tube containing 5.0 µl 10X NEB Buffer 2.

3.9. P2 Adapter ligation

1. This step in the protocol ligates the Paired_End-P2 adapter, a "Y" adapter with divergent ends that contains a 3´ dT overhang, onto the ends of blunt DNA fragments with 3´ dA overhangs.

2. To the eluate from previous step, add: 1.0 µl Paired_End-P2 Adapter (10 µ*M*), 0.5 µl rATP (100 m*M*), 0.5 µl concentrated T4 DNA Ligase. Incubate reaction at room temperature for 30 min.

3. Purify with QIAquick column. Elute in 52(50) µl EB.

OPTIONAL: 25 ml of the eluate was digested again with SbfI for 30 min to remove rare genomic DNA concatemers formed from re-ligation of short fragments with two SbfI restriction sites within 500 bp. The sample was purified, eluted in 50 ml and quantified

using the Quant-iT TM dsDNA HS Assay Kit and Qubit TM fluorometer (Invitrogen).

3.10. RAD tag Amplification/Enrichment

1. In this step you will perform high-fidelity PCR amplification on P1 and P2 adapter-ligated DNA fragments, enriching for RAD tags that contain both a P1 and P2 and preparing them to be hybridized to an Illumina Genome Analyzer flow cell (*see* **Fig. 1**).

2. **OPTIONAL?:** Perform a test amplification to determine library quality. In thin-walled PCR tube, combine: 10.5 µl H_2O, 12.5 µl Phusion High-Fidelity Master Mix, 1.0 µl Solexa primer mix (10 µ*M*), 1.0 µl RAD library template (eluate from last step). Perform 18 (14) cycles of amplification in thermal cycler: 30 sec 98° C, 18X [10 sec 98° C, 30 sec 65° C, 30 sec 72° C], 5 min 72° C, hold 4° C. Run 5.0 µl PCR product in 1X Orange Loading Dye out on 1.0% agarose gel next to 1.0 µl RAD library template and 2.0 µl GeneRuler 100 bp DNA Ladder Plus (**Fig. 3**).

3. If the amplified product is at least twice as bright as the template, perform a larger volume amplification (typically 50-100 µl) to create enough to retrieve a large amount of the RAD tag library from the final gel extraction in the protocol.

(NOTE AND SUGGESTION: (circa 40 ng is used as template in a 100 µl PCR): some authors carry out the final amplification in two separate 50 µL PCRs per library each (25 µl Phusion High-Fidelity Master Mix, 2.5 µl Solexa primer mix (10 µ*M*), 2.5 µl RAD library template

(eluate from last step), 20 µl H_2O. The two aliquots (2 x 50 µL) are combined before the final size selection).

If amplification looks poor, use more library template in a second test PCR reaction (*see* **Note 9**). **Fig. 3** shows three libraries that amplified well, which is apparent when comparing the amplified product to the amount of template loaded in the lane to the right of each sample. Template should be dim, yet visible on the gel. Purify large volume reaction with a MinElute coulumn. Elute in 20 µl EB.

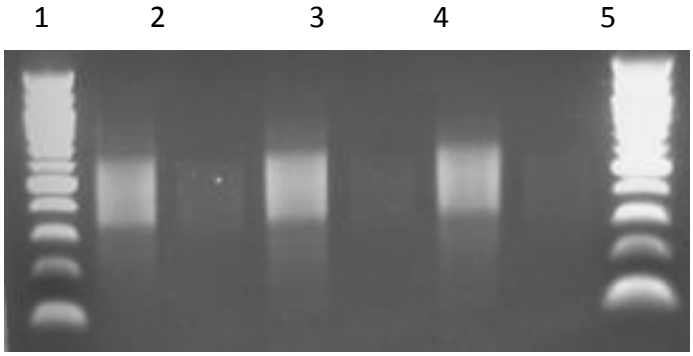


Fig. 3. Test amplification PCR product from the three libraries shown in **Fig. 2**. **2**, **4** & **6** contain 5.0 µl amplified PCR product. **2** - F_0 *Sbf*I library. **4** - F_2 *Sbf*I library. **6** - F_0 *EcoR*I library. **3**, **5** & **7** contain 1.0 µl template used for amplification in the lane to the left. Template was loaded at 5X the amount used in the equivalent volume loaded for amplified reactions. **1** & **8**

4. This purification step is performed to eliminate any contaminant bands that may appear due to an improper ratio of P1 adapter to restriction-site compatible ends (*see* **Note 6**). Load entire sample in 1X Orange Loading Dye on a 1.25% agarose, 0.5X TBE gel and run for 45 min at 100 V, next to 2.0 µl GeneRuler 100 bp DNA Ladder Plus for size reference (**Fig. 4**). Being careful to exclude any free adapters or P1 dimer contaminants running at ~130 bp and below, use a fresh razor blade to cut a slice of the gel spanning 300-500(850) bp in size in an inverted triangle shape. PCR amplification of a wide-range of fragment sizes often results in biased representation of amplified products with an increased number of short fragments. We found this to be true in our current protocol, but reduced the effects by selecting a triangular slice during gel extraction to reduce the level of short fragment lengths

from the PCR reaction. Extract DNA using MinElute Gel Purification Kit following manufacturer's instructions. Melt agarose gel slices in the supplied buffer at room temperature. Elute in 20 µl EB. (optional: diluted to 10 nM).

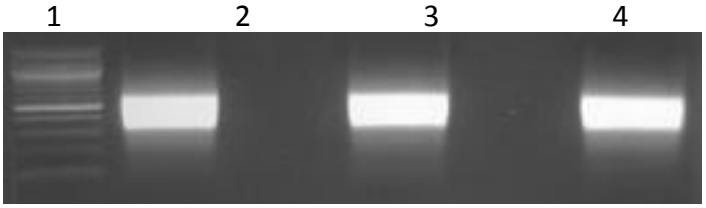


Fig. 4. PCR product from the three libraries shown in **Figs. 2 & 3** after the final large volume amplification and purification. **2**, **4** & **6** each contain 20 µl purified PCR product from 100 µl amplifications. **2** - F_0 *Sbf*I library. **4** - F_2 *Sbf*I library. **6** - F_0 *EcoR*I library. **1** - 2.0 µl

5. Quantify the DNA using a fluorimeter to get the most accurate concentration readings. Concentrations will range from 1-20 ng/µl. Determine the molar concentration of the library by examining the gel image and estimating the median size of the library smear, which should be around 400 bp. Multiply this size by 650 (the molecular mass of a base-pair) to get the molecular weight of the library. Use this number to calculate the molar concentration of the library (*see* **Note 10**).

Sequenced on the Paired-end module of the Genome Analyzer following Illumina protocols for 2x100 bp reads.

6. **IGNORE**: Validate library by cloning 1.0 µl of the gel purified library into a blunt-end compatible sequencing vector. Sequence individual clones by conventional Sanger sequencing. Verify that the insert sequences are from the genomic source DNA.

7. Sequence libraries on Illumina Genome Analyzer following manufacturer's instructions.

4. Notes

1. This protocol has been modified from that used in Baird et al., 2008 ***(12)*** and now incorporates critical improvements we have made since publication, including ones adopted from Quail et al., 2008 ***(14)***. Although we recommend following the described protocol exactly as stated, using the reagents we suggest, competing companies may offer cheaper

versions or reagents that come at lower enzyme concentrations that will work just as well. Use of these reagents may require additional optimization, including increased incubation time or larger reaction volumes, for optimal RAD tag library preparation. For instance, QIAquick columns may be substituted for MinElute columns in many places; however, reaction volumes in the subsequent step will have to be increased because of the increased elution volumes required for maximum recovery from the QIAquick columns.

2. Unless stated otherwise, all solutions should be prepared in water that has a resistivity of 18.2 MΩ-cm and total organic content of less than five parts per billion. This standard is referred to as "water" or "H_2O" in this text.

3. We recommend using a fluorescence-based method for DNA quantification in order to get the most accurate concentration readings. Since they bind specifically to double-stranded DNA, the dyes used in fluorimetric assays are not as affected by RNA, free nucleotides or other contaminants commonly found in DNA preparations (which can lead to inaccurate concentration predictions when using absorbance). If using another form of DNA quantification, such as UV spectrometer 260/280 absorbance readings, be sure to confirm the concentration by running a sample on an agarose gel and comparing to a known quantity of DNA or ladder. We recommend checking the integrity of at least a subset of samples on a gel prior to embarking on this protocol regardless of the quantification method, especially when working with many samples. Genomic DNA should consist of a fairly tight high molecular weight band without any visible degradation products or smears. When working with degraded DNA samples is the only option we have found that parameters of the protocol can be optimized (such as using more input DNA to start with and shearing less) to create usable libraries. These libraries often don't amplify as well as ones made with intact, high-quality genomic DNA.

4. Prepare 100 µ*M* stocks for each single stranded oligonucleotide in 1X Elution Buffer (EB: 10m*M* Tris-Cl, pH 8.5). Combine complementary adapter oligos at 10 µ*M* in 1X AB (10X AB: 500 m*M* NaCl, 100 m*M* Tris-Cl, pH 7.5-8.0). Place in beaker of water just off the boil, cool slowly to room temperature to anneal. Dilute to desired concentration in 1X AB. The presence of some salt is necessary for the double-stranded adapters used in this protocol to hybridize and to remain stable at ambient temperatures and above. At a 1m*M* salt

concentration the P1 adapter, which has 62 bases of complementary double-stranded sequence (assuming a 4 base pair barcode), has a T_m of approximately 40º C (depending on the barcode composition). P2, which has only 24 complementary bases, has a T_m of only 27º C at the same salt concentration. At 50 m*M* salt the T_ms jump up to ~69º and 56º, respectively.

5. In general, making master mixes, using multi-channel pipettes and dealing with samples in 96- or 384-well plates will speed up the restriction digest and P1 ligation steps when multiplexing multiple barcoded individuals.

6. *EcoR*I has been shown to work robustly in multiple organisms in our lab. **Restriction enzymes that cut less frequently create fewer RAD tags, and thus require more input DNA and less P1 adapter to keep the molar ratio approximately equal.** Less frequent cutters are more difficult to amplify in general and protocol parameters may take some optimization for favorable results. It is critical to optimize the amount of P1 adapter used when a given restriction enzyme is used for the first time in an organism, unless the actual number of sites is known. Otherwise, some optimization may be required to ensure enough P1 is used to get robust RAD library amplification without using too much. If the ratio of P1 adapter overhangs to available genomic compatible ends is too low, you will get insufficient amplification and/or biased representation of some RAD tags. However, if the ratio of P1 to genomic overhangs is too high, **a contaminant band that runs around 130 bp will appear after the final PCR reaction**. If this contaminant overwhelms the amplification reaction it can lead to significant adapter sequence reads in the final sequencing output (even after gel extraction following the final PCR). This phenomenon is completely dependent upon the number of actual cut sites present in that genome. Our *Sbf*I study in stickleback used 2.5 µl P1 per microgram starting material and performed very well for library construction (*see* **Figs. 3** & **4**; lanes 2 & 4); however, this is likely to due to the fact that there are actually more *Sbf*I sites than expected by chance. Therefore, starting with less P1 may be preferable for genomes with closer to the expected number of sites.

ligation molarity calculator - guide for calculating appropriate adapter concentration for ligations.

7. DNA samples from 96 recombinant F_2 individuals were uniquely barcoded, which allowed us to track RAD markers and associate them with differing lateral plate or pelvic structure phenotypes. F_2 individuals used in the mapping analysis included 60 fish possessing the complete lateral plate phenotype, 31 low lateral plate individuals. The barcoded samples from fish possessing the same lateral plate phenotype were combined and treated as one library after P1 ligation. In order to genotype all F_2 individuals with a low pelvic structure phenotype, the DNA from 5 individuals that had a low pelvic score, but that had a partial lateral plate phenotype, were barcoded and processed with the low plate group. The two multiplexed libraries included 67 individuals demonstrating the high pelvic structure phenotype and 29 with a low pelvic score that were resorted *in silico* to map this second trait. For the bulk-segregant analysis using *Sbf*I, one library was prepared using two pooled DNA samples from recombinant F_2 individuals, combined according to lateral plate phenotype prior to restriction digestion. The digested pools were labeled with different barcodes, combined and treated as one library after P1 ligation.

8. We have found it is unwise to run more than one library sample on the same agarose gel, as is shown in the figures, unless they will be combined and sequenced in the same lane on the flow cell, because it can lead to contamination between samples. This is especially important when dealing with samples following PCR amplification. We recommend using aerosol-resistant filter tips for all amplification and downstream steps in the protocol to avoid library contamination.

9. Libraries that amplify robustly, such as those shown in **Fig. 3**, can be amplified with only 16 or fewer cycles of PCR to avoid skewing the representation of the library ***(14)***. Robust libraries can often times be cleaned up without the final gel extraction step if there are no visible contaminant bands running below 300 bp on the gel after the test amplification. We have retrieved good sequence read numbers from libraries that amplified well with only 4 µl of template in a 100 µl reaction as well as ones that amplified very poorly and required up to 30 µl template in the same volume. The first scenario is preferable, as you can be more confident of the true concentration of RAD tag molecules in your sample, which have both P1 and P2 sequences, and are therefore able to bind to adapter oligonucleotides present on the Illumina flow cell. Poorly amplified libraries will contain a greater number of background

sheared genomic DNA fragments with only P2 adapters attached, which cannot bind to the flow cell.

10. For long-term storage of DNA samples, Illumina recommends a concentration of 10 n*M* and adding Tween 20 to the sample to a final concentration of 0.1% Tween. This helps to prevent adsorption of the template to plastic tubes upon repeated freeze-thaw cycles, which would decrease the cluster numbers from a sample over time.

References

1. Berger J, Suzuki T, Senti KA, Stubbs J, Schaffner G et al. (2001) Genetic mapping with SNP markers in Drosophila. Nat Genet 29: 475-481.

2. Stickney HL, Schmutz J, Woods IG, Holtzer CC, Dickson MC et al. (2002) Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. Genome Res 12: 1929-1934.

3. Wicks SR, Yeh RT, Gish WR, Waterston RH, Plasterk RH (2001) Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. Nat Genet 28: 160-164.

4. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E et al. (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A 101: 9915-9920.

5. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-331.

6. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23: 4407-4414.

7. Chen D, Ahlford A, Schnorrer F, Kalchhauser I, Fellner M et al. (2008) High-resolution, high-throughput SNP mapping in Drosophila melanogaster. Nat Methods 5: 323-329.

8. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17: 240-248.

9. van Orsouw NJ, Hogers RC, Janssen A, Yalcin F, Snoeijers S et al. (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 2: e1172.

10. Miller MR, Atwood TS, Eames BF, Eberhart JK, Yan YL et al. (2007) RAD marker microarrays enable rapid mapping of zebrafish mutations. Genome Biol 8: R105.

11. Lewis ZA, Shiver AL, Stiffler N, Miller MR, Johnson EA et al. (2007) High-density detection of restriction-site-associated DNA markers for rapid mapping of mutated loci in Neurospora. Genetics 177: 1163-1171.

12. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE. 3(10):e3376. Epub Oct 13.

13. Coyne KJ, Burkholder JM, Feldman RA, Hutchins DA, Cary SC (2004) Modified serial analysis of gene expression method for construction of gene expression profiles of microbial eukaryotic species. Appl Environ Microbiol 70: 5298-5304.

14. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods. Dec;5(12):1005-10.